

KOS.content

01 | 2015

Ergebnisse der Untersuchungen des
Kompetenzzentrum Open Source der DHBW-Stuttgart

Frühjahr 2015
band.2

INHALT BAND.2

Inhalt __

Das Kompetenzzentrum Open Source (KOS)

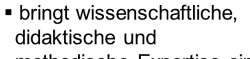
Ziel des Projektes

Das Projekt Kompetenzzentrum Open Source der DHBW Stuttgart wurde mit der Zielsetzung ins Leben gerufen, die Einsatzfelder für Open Source Software in Unternehmen zu identifizieren und durch den Einsatz quelloffener Produkte und deren kostengünstigen Einsatzmöglichkeiten Optimierungen in ausgewählten Geschäftsbereichen zu erzielen.

Dies bedeutet konkret, dass z.B. Open Source Software evaluiert wird, um Lizenzkosten zu reduzieren, bewertet wird, ob sie diverse Qualitätskriterien erfüllt und erfolgreich(er) und effizient(er) in Unternehmen genutzt werden kann. Das Ziel des Projektes ist es hierbei, allgemeingültige Lösungskonzepte für Problemstellungen zu erarbeiten, welche von den am Projekt beteiligten Unternehmen zu firmenspezifischen Lösungen weiterentwickelt werden können. Die beteiligten Unternehmen partizipieren so an den Ergebnissen des Projekts.

Zusammenarbeit mit den Dualen Partnern

Die Zusammenarbeit mit den Dualen Partnern gestaltet sich entlang deren Anforderungen und Bedürfnissen. Sie sind die Themengeber für betriebliche Fragestellungen, die im Rahmen des Projekts untersucht werden. Die DHBW steuert die wissenschaftliche, didaktische und methodische Expertise und Forschungskompetenz bei und untersucht die identifizierten Themenfelder.






Im Rahmen des Projektes steuert die DHBW Stuttgart die wissenschaftliche Expertise und Forschungskompetenz bei zur Bearbeitung der betrieblichen Fragestellungen der Dualen Partner. Es entstehen generische Lösungen, welche von den Partnern an Ihre Situation angepasst werden kann.
Im Rahmen der Arbeit entstehen (generische) Lösungen, an denen die Partner teilhaben können indem sie diese auf ihre spezifische Unternehmenssituation anpassen. Zudem fließen die Ergebnisse in die Arbeit der DHBW ein, sodass hier dem Anspruch an eine hohe Anwendungs- und Transferorientierung ganz im Sinne einer kooperativen Forschung Rechnung getragen wird.

An den Ergebnissen des Projekts partizipieren die Dualen Partner Allianz Deutschland AG, die Deutsche Rentenversicherung Baden-Württemberg, die HALLESCHE Krankenversicherung a.G. und die WGV-Informatik und Media GmbH.



Testszenarien für NoSQL-Datenbanksysteme/-dienste aus der Cloud

Schriftliche Ausarbeitung
im Rahmen der Lehrveranstaltung „Integrationsseminar“
für das Kompetenzzentrum Open Source (KOS)

Vorgelegt von

Eitel, Fabian
Neef, Jacqueline
Seibold, Marcel
Weber, Maximilian

am 26.01.2015

Fakultät Wirtschaft
Studiengang Wirtschaftsinformatik
WI2012I

Inhaltsverzeichnis

Abkürzungsverzeichnis

ASF	Apache Software Foundation
AWS	Amazon Web Services
DBMS	Datenbankmanagementsystem
DBaaS	Database as a Service
EMR	Elastic MapReduce
ENISA	European Network and Information Security Agency
IBM	International Business Machines
IaaS	Infrastructure as a Service
NIST	National Institute of Standards and Technology
PaaS	Plattform as a Service
SaaS	Software as a Service
URI	Uniform Resource Identifier

Abbildungsverzeichnis

1 Einleitung

1.1 Problemstellung und Zielsetzung

Wissen ist Macht. Das gilt insbesondere für heutige Unternehmen. Folglich werden Daten immer wichtiger für die moderne Wirtschaft.[1] Insbesondere da diese sich als Entscheidungsgrundlage anbieten. Die große Herausforderung ist dabei die Vielzahl der Daten schnell und effizient aufzubereiten. Bei großen Mengen an unstrukturierten Daten kommen klassische, relationale Datenbanksysteme schnell an ihre Grenzen. Ein alternativer Lösungsansatz sind NoSQL-Datenbanksysteme zur effizienten Speicherung der Daten – die vermehrt als Dienstleistung in der Cloud angeboten werden. Für Unternehmen besteht dabei die große Frage, für welche Testszenarien sich eine solche Datenbank aus der Cloud anbietet.

Im Rahmen dieser Arbeit wird ein NoSQL-Datenbanksystem und der jeweilige Anbieter aus der Cloud vorgestellt. Anschließend gilt es, passende Testszenarien zu konzipieren und praktisch umzusetzen. Abschließend wird eine Aussage über die Komplexität der Umsetzung und der Höhe der Kosten getroffen.

1.2 Struktur und Beiträge der Arbeit

Zunächst werden in Kapitel 2 die technischen Grundlagen der eingesetzten NoSQL-Datenbanksysteme sowie das Hadoop-Framewok und das Datenbankmanagementsystem CouchDB skizziert. Im Folgenden die Servicemodelle der Cloud-Computing-Anbieter kategorisiert.

Daraufhin erfolgt in Kapitel 3 die Recherche, Auswahl und Analyse potenzieller Datensätze, die den Anforderungen eines „Big-Data-Datensatzes“ basierend auf einem selbst erstellten Anforderungs-Framework standhalten, frei verfügbar sind oder im Zweifelsfall aggregiert werden müssen. Zusätzlich werden in diesem Kapitel zwei verschiedene Testszenarien auf der Amazon Cloud Plattform „Amazon Web Services“ sowie auf IBM Cloudant implementiert. Dabei werden neben der reinen Vorgehensweise auch die Ergebnisse der Analysen festgehalten.

Abschließend wird im Fazit der Einsatz von NoSQL-Datenbanksystemen bewertet und kritisch beleuchtet sowie ein Ausblick für die praktische Verwendung von NoSQL-Datenbanksystemen aus der Cloud gegeben.

[1] Vgl. Weber, M. (2012), S. 11

2 Theorie

2.1 NoSQL im Überblick

Der Begriff „NoSQL“ wurde erstmals 1998 von Carlo Strozzi verwendet, um sein relationales Datenbanksystem, welches nicht die Datenbanksprache SQL verwendet, zu beschreiben.[2] Seit 2009 wird mit NoSQL zunehmend die Bedeutung *„Not only SQL“* assoziiert.[3] Das ist darauf zurückzuführen, dass die meisten NoSQL-Datenbanken die Eigenschaft haben, auf Schemata zu verzichten – im Gegenteil zu den relationalen Pendants. Mithilfe flexibler Techniken definieren sie, wie Daten gespeichert werden, oder überlassen dies der Anwendung.[4] NoSQL-Datenbanken dürfen nicht als das Gegenteil von relationalen Datenbankmodellen verstanden werden. Vielmehr handelt es sich hier um eine alternative Weiterentwicklung, die versucht deren Nachteile auszumerzen. Grundsätzlich gehört zu dem Begriff „NoSQL“ eine Vielzahl heterogener Datenbanksysteme, die sich teilweise deutlich in ihrer Architektur, ihrer Funktion und ihrem Einsatzfeld unterscheiden.[5]

Die Einsatzmöglichkeiten von NoSQL-Datenbanken sind vielseitig. Besonders Sinn machen sie, wo die Leistung von klassischen SQL-Datenbanken nicht mehr ausreicht oder die Umsetzung mit einer sehr komplexen Architektur verbunden ist. Das trifft zu, wenn große Datenmengen angefragt und verarbeitet werden müssen. Typische Praxisbeispiele sind Streamingdienste wie Netflix, die über NoSQL-Datenbanken großen Mengen an hochauflösendem Videomaterial ohne an Leistungsgrenzen zu stoßen, verarbeiten können.[6]

2.2 Theoretische Konzepte der NoSQL-Bewegung

Im Rahmen dieser Arbeit werden einige dieser teilweise recht unterschiedlichen Technologien der NoSQL-Bewegung genauer betrachtet. Dabei wird der Fokus auf jenen Konzepten liegen, welche innerhalb der Testszenarien ihre Anwendung finden. Es werden Apache Hadoop als Softwareframework, das Map-Reduce Verfahren als Programmiermodell und CouchDB als dokumentenorientiertes DBMS aus der Cloud, vorgestellt.

[2] Vgl. Kuznetsov, S./ Poskonin, A. (2014), S.1
[3] Vgl. Edlich, S. (2015)
[4] Vgl. Walker-Morgan, D.(2010)
[5] Vgl. Klimt, W. (2013)
[6] Vgl. Izrailevsky, Y. (2011)

2.2.1 Apache Hadoop

Apache Hadoop wurde 2005 im Rahmen eines Projekts der Apache Software Foundation (ASF) ins Leben gerufen. Es ist ein in Java programmiertes Open-Source-System zur Verarbeitung sehr großer Datenmengen. Traditionell stütze sich das Hadoop Projekt auf eine 2004 von Google Inc. veröffentliche Arbeit zum MapReduce-Verfahren und wurde durch das Google File System beeinflusst.[7]

Historisch ist Hadoop aufgrund der Tatsache entstanden, dass ein einzelner Rechner zur Verarbeitung und Analyse großer Datenmengen – sprich Big Data – technisch nicht mehr ausreicht. Um dieses Problem möglichst wirtschaftlich zu lösen, kam das Konzept der parallelen Verarbeitung auf.[8] Die Grundidee besteht darin, dass viele mäßig rechen- und speicherstarke Computer als Knoten in einem Cluster an der gleichzeitigen Verarbeitung großer Datenmengen arbeiten. Klassisch hat einer dieser Knoten eine Speicherkapazität von 2-4 TB.[9] Um die Performanz des Systems zu maximieren, werden die zu verarbeitenden Datenmengen auf die vorhandenen Knoten verteilt, sodass bei einer Verteilung auf zehn Knoten eine bis zu zehnfache Performanzsteigerung erreichbar ist.[10]

Insgesamt ist zu sagen, dass Hadoop ein in Java programmiertes System zur Verarbeitung großer Datenmengen innerhalb eines verteilten Systems ist, das unter einer Apache Lizenz verfügbar ist. Seine Funktionsweise ist so ausgelegt, dass (kleinere) Fehler und Ausfälle im Cluster aufgrund seiner automatischen Daten- und Hardware-Redundanz keine Auswirkungen haben.

Von der ursprünglichen Funktionalität von Hadoop als MapReduce ausführendes System, zur Verarbeitung großer Mengen an Textdateien, hat sich Apache Hadoop zu einem umfangreichen Modell entwickelt, das zahlreiche Use Cases in Unternehmen bedienen kann.[11] Heute stellt die ASF unter der Hadoop Plattform nicht nur das marktführende MapReduce Framework und das Hadoop Distributed File System (HDFS) zur Verfügung, sondern überdies eine Reihe weiterer Werkzeuge zur Verarbeitung und Analyse von Big Data. Hierzu zählen das Data Warehouse „Hive“, das nicht-relationale DBMS „HBase“ sowie das Hadoop Monitoring-System „Ambari“. Zu den prominenten Nutzern der Hadoop Plattform gehören namhafte Firmen wie Facebook, IBM oder auch Yahoo.[12]

[7] Vgl. Frampton, M. (2015), S. 4
[8] Vgl. Wadkar, S./ Siddalingaiah, M. (2014), S. 1
[9] Vgl. ebenda, S. 2
[10] Vgl. ebenda, S. 2
[11] Vgl. ebenda, S.12
[12] Vgl. Redmond E./Wilson J. (2012), S. 103-105

2.2.2 MapReduce-Verfahren

Oftmals gemeinhin als „Herz von Hadoop“ bezeichnet, ist das MapReduce-Verfahren zu nennen.[13] Konkret ist MapReduce ein von Google Inc. entwickeltes Modell, welches es ermöglicht parallele Berechnungen auf mehreren Petabyte großen Datensätzen innerhalb eines Rechnerverbunds durchzuführen.[14]

Das generelle Konzept zeichnet sich dadurch aus, dass unstrukturierte Daten in „logische Stücke“ aufgeteilt werden und so parallel in einem Cluster verarbeitet werden können. Das Verfahren wurde basierend auf Funktionen der funktionalen Programmierung entwickelt und zeichnet sich durch zwei wesentliche Schritte aus: Map und Reduce. Die Funktionsweise ist in Abbildung 1 grafisch dargestellt.

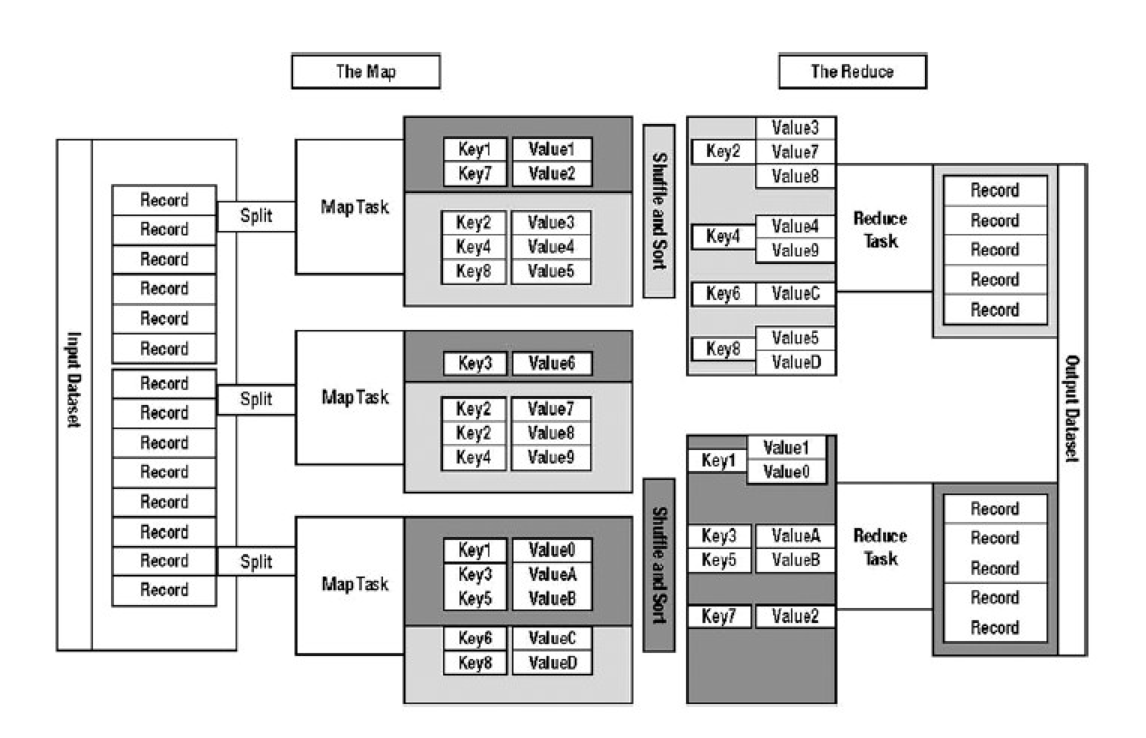


Abb.1: MapReduce Model[15]

Zunächst werden Im Map-Vorgang die unstrukturierten Daten auf verschiedene „Map Tasks“ innerhalb des Clusters verteilt. Die Map Tasks können auf jedem beliebigen Knoten ausgeführt werden; auch die parallele Verarbeitung ist möglich. Innerhalb des Map-Vorgangs werden die Daten in Key-Value-Paare umgewandelt, dann aufgeteilt, sortiert und in einem Zwischenspeicher abgelegt. Schließlich wird jedes dieser gewonnenen Datenpakete bestehend

[13] Vgl. IBM (2015)

[14] Vgl. Wadkar, S./ Siddalingaiah, M. (2014), S. 12

[15] Enthalten in: ebenda, S.13

aus Schlüsseln mit den dazugehörigen Werten einem „Reduce-Task“ unterzogen. Diese können wiederum parallel ablaufen und haben zum Ziel die Daten dergestalt zusammenzufassen wie sie der Nutzer benötigt.[16] Die Tatsache, dass der Nutzer des MapReduce-Verfahrens entscheiden kann, mittels welcher Logik die Daten zunächst in der Map-Phase und schließlich in der Reduce-Phase verarbeitet werden, macht das MapReduce-Verfahren sehr flexibel. Jedoch muss der Nutzer ausschließlich die beiden getrennt voneinander zu betrachtenden Prozesse definieren: map und reduce. Die Datenverteilung auf die Knoten, sowie die Verwaltung der Zwischenspeicher geschieht automatisiert.[17]

2.3 CouchDB

Ein populärer Vertreter der NoSQL-Datenbanksysteme ist Apache CouchDB. Gemäß seinem Slogan „Apache CouchDB has started. Time to relax“[18] hat es die Zielsetzung mittels einfacher Grundkonzepte auch für Anfänger leicht zu bedienen zu sein. Bei Apache CouchDB handelt es sich um ein dokumentenorientiertes Datenbanksystem, das als Open-Source-Produkt unter der Apache 2.0-Lizenz verfügbar ist.[19] Das Akronym CouchDB steht für „Cluster Of Unreliable Commodity Hardware Data Base“.[20]

Das verdeutlicht die Eigenschaften von CouchDB: Hohe Skalierbarkeit, Verfügbarkeit und Zuverlässigkeit – auch wenn es auf Hardware läuft, die fehleranfällig ist.[21] Historisch betrachtet, beginnt die Geschichte von CouchDB mit Lotus Notes. Lotus Notes wurde in den 80er-Jahren entwickelt und ist eines der ersten dokumentenorientierten Datenbanksysteme[22].

Im Jahre 1995 wurde es von IBM übernommen und in IBM Notes umbenannt. Der in dieser Zeit bei IBM tätige Damien Katz entwickelte dann in 2005 CouchDB.[23] Seit 2008 übernimmt die Apache Software Foundation die Entwicklung von CouchDB, weshalb es heute den Namen Apache CouchDB trägt. „Couch is Lotus Notes built from the ground up for the Web“[24]. Dieses Zitat des Erfinders von CouchDB, Damien Katz, verdeutlicht, dass der Einsatzbereich des Datenbanksystems vor allem bei Webanwendungen liegt. Dazu passend kann zur komfortablen Verwaltung von CouchDB die integrierte Web-Oberfläche namens „Futon“ verwendet werden.

[16] Vgl. Wadkar, S./ Siddalingaiah, M. (2014), S. 14f
[17] Vgl. ebenda, S. 5
[18] Vgl. The Apache Software Foundation (2014)
[19] Vgl. Wenk, A. (2014)
[20] Vgl. Lennon, J. (2009)
[21] Vgl. ebenda
[22] Vgl. IBM (2007)
[23] Vgl. Jansen, R. (2010)
[24] Vgl. Katz, D. (2005)

Die Funktionsweise von Apache CouchDB basiert auf dem Paradigma der dokumentenorientierten Speicherung.[25] Die Daten werden nicht in Form von Tabellen gespeichert, wie das bei relationalen Datenbanken der Fall ist. Stattdessen werden sie als Dokumente in der Datenbank abgelegt. Ein Dokument wird als eine strukturierte Zusammenstellung bestimmter Daten verstanden.[26] Im Fall von CouchDB werden die Daten als JSON-Dokumente gespeichert.

```
{
"Vorname" : "Max" ,
"Nachname" : "Mustermann" ,
"Telefon-Nr." : "0123456",
"Adresse" : "Musterstraße 34, Musterstadt" ,
"Kinder" :  ["Musterkind1" , "Musterkind2" ] ,
"Alter" :  33
...
}
```

Abb.2: Beispiel eines Dokuments im JSON-Format[27]

Wie ein solches Dokument im JSON-Format aussehen kann, zeigt Abbildung 2. Die Daten innerhalb des Dokuments werden als sogenannte Schlüssel/Wert-Paare (Key/Value) gespeichert. Beispielsweise ist in der Abbildung ein Schlüssel „Vorname“ und der dazugehörige Wert „Max“. Über diese Schlüsselfelder („Nachname“, „Telefon-Nr.“, „Kinder“) kann auf die jeweiligen Werte im Dokument zugegriffen werden. Jegliche Information kann als Wert abgespeichert werden – unabhängig davon, ob es sich um Wörter, Nummern, Listen oder Maps handelt. Eine Datenbank basierend auf CouchDB ist eine flache Ansammlung solcher Dokumente. Jedes Dokument ist mit einer eindeutigen ID versehen.[28]

CouchDB wurde mit dem Ziel entwickelt, große Mengen an semi-strukturierten, dokumentenbasierten Dokumenten abspeichern und verarbeiten zu können. Neue Dokumente, die

[25] Vgl. Rudolf, J. (2010)
[26] Vgl. Datenbanken Online Lexikon (2013a)
[27] Vgl. ebenda
[28] Vgl. Datenbanken Online Lexikon (2013b)

eine neue Bedeutung haben, können einfach in der CouchDB Datenbank abgespeichert werden. Folglich ist es nicht erforderlich im Voraus festzulegen, welche Daten in der Datenbank gespeichert werden können und in welcher Beziehung diese zueinanderzustehen haben. Diese bei relationalen Datenbanken notwendige Schemadefinition entfällt bei CouchDB.[29] Das bedeutet aber auch, dass Überprüfungs- und Aggregationsfunktionen sowie Beziehungen zwischen den Dokumenten mit Hilfe von Views eigenhändig implementiert werden müssen.[30] Ein View ist eine Teilmenge der Datenbank, die das Ergebnis einer Abfrage enthält – in Form eines permanenten Objekts.[31] Apache CouchDB nutzt das Map/Reduce-Verfahren von Google, um mittels Computerclustern große Datenmengen verarbeiten und generieren zu können.

Die Kommunikation mit CouchDB findet über HTTP-Anfragen statt, was die Anbindung an Webanwendungen enorm vereinfacht. Zusätzlich muss hierdurch kein eigenes Kommunikationsprotokoll definiert werden.[32] Als Skriptsprache verwendet CouchDB JavaScript, was serverseitig von SpiderMonkey interpretiert wird.[33]

Um konkurrierende Zugriffe auf die Datenbanken zu verwalten, verwendet ApacheCouchDB das Prinzip der Multi-Version Concurrency Control (MVCC). Anfragen werden parallel beantwortet, wodurch das Datenbanksystem selbst unter hoher Last mit voller Geschwindigkeit arbeitet. Jede Änderung in einem Dokument führt zu einer komplett neuen Version dieser, die zusätzlich abgelegt wird. Dadurch ist der parallele Zugriff immer gewährleistet. Während eine erste Anfrage ein Dokument ausliest, kann die zweite Anfrage zeitgleich das Dokument ändern. Eine neue Version dieses Dokuments wird an der Datenbank angeknüpft. Auf das Ende der ersten Anfrage muss nicht gewartet werden. Wenn das Dokument innerhalb einer dritten Anfrage ausgelesen wird, wird die neue Version ausgegeben. Während dieser Zeit kann die erste Anfrage immer noch die ursprüngliche Version auslesen. Bei einem Lesezugriff wird also immer der aktuellste Zustand der Datenbank verwendet.[34]

Wie eingangs bereits angesprochen, ist CouchDB vor allem für die Verwendung bei Webanwendungen geeignet. Die Eingabe beziehungsweise Speicherung von Dokumenten ist durch den Wegfall von Schemadefinitionen sehr komfortabel gestaltet. Auch bei einer Vielzahl von Anfragen kann es dank der MVCC in voller Geschwindigkeit arbeiten und Anfragen parallel beantworten. Eine Herausforderung sind komplexe Anfragen, da diese selbstständig implementiert werden müssen. Zusammenfassend differenziert sich CouchDB vor allem durch die

[29] Vgl. Datenbanken Online Lexikon (2013b)
[30] Vgl. Lennon, J. (2009)
[31]Vgl. Janssen, C. (o.J.)
[32] Vgl. Datenbanken Online Lexikon (2013b)
[33] Vgl. ebenda
[34] Vgl. Anderson, C. / Lehnardt, J. / Slater, N. (2015)

Speicherung von Daten als JSON-Dokumente, die Verwendung von HTTP zur Kommunikation, der Zuverlässigkeit sowie durch die Konsistenz der Datenspeicherung.[35]

2.4 Cloud Provider

Zahlreiche NoSQL-Datenbanksysteme sind heute aus der Cloud, also via Cloud Computing, verfügbar. Folglich ist es für das Verständnis dieser wissenschaftlichen Arbeit bedeutend, den Begriff Cloud Computing zu definieren. In der Fachliteratur wird meist auf die Definition des National Insitute of Standards and Technology (NIST), der US-amerikanischen Standardisierungsstelle, verwiesen:

„Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.“[36]

Diese Definition unterstreicht dabei die Vorzüge von Cloud Computing und insbesondere den minimalen Verwaltungsaufwand, beziehungsweise den geringen Bedarf an Interaktionen mit dem Service Provider. Kritisch anzumerken ist, dass hier eine sehr technische Sichtweise für die Definition herangezogen wird. Cloud Computing ist ein Thema, das aber außerhalb der technischen Abteilungen ebenfalls an großer Bedeutung gewonnen hat. Folglich sollte die wirtschaftliche Sicht nicht vernachlässigt werden. Eine Definition, die diese Voraussetzung erfüllt, stammt vom Bundesamt für Sicherheit in der Informationstechnik (BSI):

„Cloud Computing bezeichnet das dynamisch an den Bedarf angepasste Anbieten, Nutzen und Abrechnen von IT-Dienstleistungen über ein Netz. Angebot und Nutzung dieser Dienstleistungen erfolgen dabei ausschließlich über definierte technische Schnittstellen und Protokolle. Die Spannbreite der im Rahmen von Cloud Computing angebotenen Dienstleistungen umfasst das komplette Spektrum der Informationstechnik und beinhaltet unter anderem Infrastruktur (z. B. Rechenleistung, Speicherplatz), Plattformen und Software.“[37]

Hier wird deutlich hervorgehoben, dass Cloud Computing das komplette Spektrum der IT abdecken kann. Ebenfalls wird erwähnt, dass es im Hinblick auf Cloud Computing drei verschiedene Kategorien an Dienstleistungen gibt.

[35] Vgl. Datenbank Online Lexikon (2013b)
[36] National Institute of Standards and Technology (2011), S. 2
[37] Bundesamt für Sicherheit in der Informationstechnik (o.J.)

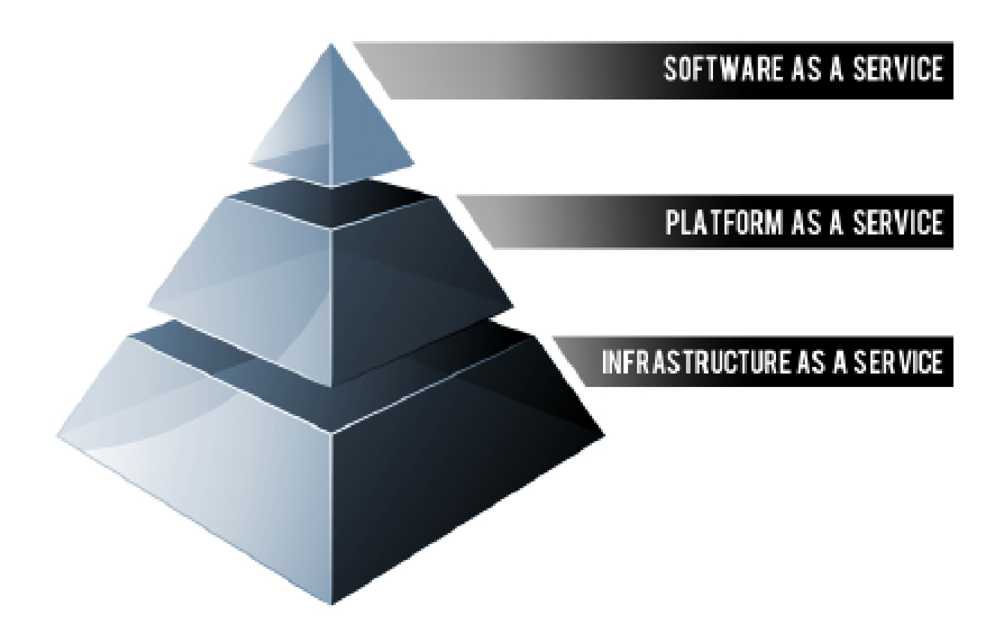


Abb.3: Kategorien der Cloud Services[38]

Diese sind in Abbildung 3 gemäß ihrer Ebene in der IT gestaffelt: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) und Software as a Service (SaaS). Diese drei unterschiedlichen Service Modelle sind für das Verständnis von NoSQL-Datenbanken aus der Cloud bedeutend und werden daher kurz erläutert.

Bei Infrastructure as a Service (IaaS) werden grundlegende IT Ressourcen als Dienstleistung angeboten. Beispielsweise Rechenleistung, Netze oder, wie bei NoSQL-Datenbanken, Datenspeicher.[39] Der große Vorteil für den Kunden ist, dass er keine teure Anschaffungs- und Instandhaltungskosten von Hardware hat und dass er nur für seinen konkreten Bedarf zahlt.

Platform as a Service ist die Middleware des Cloud-Computings. Sie ermöglicht Software Entwicklern schnell Software zu entwickeln – ohne sich um die Installation und Wartung von Software und Hardware beschäftigen zu müssen. Die Lösungen enthalten oft vorkonfigurierte Pakete, wodurch die Verbindung mit dem IaaS-Portfolio einfacher und schneller möglich ist.[40]

Die letzte Kategorie ist Software as a Service (SaaS). Dazu zählen jegliche Applikationen, die in der Cloud ausgeführt werden und dem Kunden einen direkten Service bieten. Auf die Applikationen wird meist mit einem Webinterface zugegriffen. Updates und Wartungsaktio-

[38] Enthalten in: Rackspace Support (2013)

[39] Vgl. Bundesamt für Sicherheit in der Informationstechnik (o.J.)

[40] Vgl. u.a. Baun, C. (2011), S. 35

nen werden vom SaaS-Anbieter durchgeführt, wodurch der Endnutzer meist nicht beeinflusst wird.[41]

Im Hinblick auf die Verwendung von Datenbanken aus der Cloud wird oftmals der Begriff Database as a Service (DBaaS) verwendet. Hierbei handelt es sich um eine Dienstleistung aus der Cloud, bei der die Instanz für die Datenbank nicht vom Nutzer selbst gestartet werden muss. Die Installation und Instandhaltung der Datenbank übernimmt der Anbieter. Der Kunde zahlt entsprechend seiner Datenbanknutzung. Gemäß den Anforderungen des Kunden kann dieser natürlich auch mehr Kontrolle über seine Datenbank erhalten und die Administration bis zu einem gewissen Grad selbst verwalten.[42] Es können auch NoSQL Datenbanken im Rahmen von DBaaS genutzt werden.

Der Markt an DBaaS ist für das Jahr 2014 auf ein Volumen von $1,07 Mrd. beziffert. Bis 2019 soll er sich auf $14.05 Mrd. Erhöhen.[43] Große IT-Dienstleister wie IBM oder Amazon haben das ebenfalls erkannt und bieten schon bereits länger DBaaS-Produkte an. Im Produktportfolio der IBM befindet sich Cloudant – eine NoSQL-Datenbank, die auf CouchDB basiert und als DBaaS angeboten wird.[44] Amazon bietet gleich mehrere Cloud-Dienste an, die unter der Bezeichnung Amazon Web Services geführt werden. Dazu gehören auch DBaaS-Produkte wie Amazon DynamoDB und Amazon Elastic Map Reduce (EMR). Bei Letzterem handelt es sich um einen NoSQL-Dienst, der auf dem Hadoop Framework basiert und auf der Amazon eigenen Cloud Computing-Infrastruktur läuft.[45] Amazon stellt innerhalb dieses Webservices neben einem Cluster für die Verarbeitung von Big Data auch weitere Werkzeuge wie Simulationssoftware und Data Mining Lösungen zur Verfügung.[46]

[41] Vgl. u.a. Lenk, A. (2009), S. 2
[42] Vgl. ScaleDB (o.J.)
[43] Vgl. Clustrix, Inc. (2014)
[44] Vgl. IBM (2014), S.1
[45] Vgl. AWS (o.J.)
[46] Vgl. Amazon Web Services (2015a)

3 Praxis

3.1 Untersuchung von Testdaten

Um in der kurzen Projektzeit möglichst effizient sinnvolle Daten für die praktischen Tests auszuwählen, wurde zur Überprüfung ein Framework herangenommen, das auf Basis der Definition der Merkmale von Big-Data-Datensätzen im Rahmen dieser wissenschaftlichen Arbeit erstellt wurde. Trotz dessen gestaltete sich die Recherche nach kostenlosen sowie frei zugänglichen, sogenannten „Open-Data-Sets“ als komplexe Aufgabenstellung. Das Framework vereint verschiedene granulare Anforderungen, die nötig sind, um verschiedenen Big-Data-UseCases standzuhalten.

Werden folgende Anforderungen dieses Large-Datasets Frameworks vollständig oder teilweise erfüllt, so sind die erforderlichen Merkmale eines Big-Data Datensatzes erfüllt und es ist aus Sicht der Autoren sinnvoll, eine nicht-relationale Datenbank für die Auswertung einzusetzen.[47]

Big-Data-Merkmale	Volume	Variety	Velocity
	>500GB	Unstrukturiert	Deutlicher Datenwachstum (>100 GB in 24h)
	Fragmentierte Daten	Stetig ändernde Art auftretender Daten	Viele und komplexe Abfragen

Tabelle 1: Eigene Darstellung - Large-Datasets Framework

Öffentlich zugängliche Datensätze in ausreichender Größe sind schwer erhältlich. Bei vielen Testszenarien werden lediglich Zahlenreihen mit einem Zufallsgenerator aneinandergereiht. Im Rahmen dieses wissenschaftlichen Projektes wurde eine Reihe von Datensätzen gesichtet und nicht nur auf die schiere Größe geachtet, sondern insbesondere auf die Praktikabilität mit Anwendungsfällen aus der Wirtschaft. Dennoch sind im Internet einige kostenlose Datensätze verfügbar die eine zufriedenstellende Größe aufweisen. Eine ausführliche Recherche ergab folgende Auswahl an potenziellen Datensätzen:

Million Song Database

Die “Million Song Database” ist eine Metadaten-Datenbank zu mehreren Millionen von Musikdateien. Sie ist mit einer Gesamtgröße von 280GB jedoch relativ klein. Dabei wird unter

[47] Vgl. Weber,M. (2012), S.7

anderem der Künstlername, der Musiktitel, das Erscheinungsdatum oder das Tempo in BPM ("Beats per Minute") gespeichert. Besonders interessant ist der Datensatz deshalb, da die Track-Daten ähnliche Abfragen wie bei Personendaten ermöglichen und dabei ist auch die Größe zufriedenstellend.[48]

Wikipedia Datenbank

Auf den ersten Blick eine spannende Quelle für eine große Menge an Daten, allerdings besteht der gesamte Datenabzug der englischen Wikipedia-Seite im SQL-Format gerade einmal aus 10GB und ist damit nicht ansatzweise groß genug für eine sinnvolle Untersuchung mit einer nicht-relationalen Datenbank, enthält aber viele komplexe Informationen, die komplizierte Abfragen ermöglichen.[49]

The Human Genome Database

Als potenzieller Datensatz hat sich die "Human Genome Database" durch die schiere Größe von 200TB herausgestellt, bei der selbst simple Abfragen mit einem herkömmlichen Datenbanksystem nicht zu bewältigen wären und die Daten zusätzlich in unstrukturierter Form vorliegen, da jeweils die neusten Projektdaten hinzugefügt werden. Sie besteht aus den vollständigen Genomsequenzen von 1700 Menschen und soll bis zum Ende des Projektes aus den Sequenzen von über 2600 Menschen aus 26 Populationen bestehen.[50]

Common Crawl Database

Bedingt durch die Größe des Datensatzes ist letztendlich die Entscheidung auf die „Common Crawl Database" gefallen, die praktischerweise innerhalb der Amazon Web Services gehostet wird. Bei der Common Crawl Database handelt es sich um einen mehrere Petabyte großen Crawling-Datensatz (Rohdaten, Textdaten und Metadaten) – aggregiert aus Internet-Crawling-Daten. Der bei Amazon Web Services vorliegende Datenbestand hatte zum Zeitpunkt der Erstellung dieser Arbeit eine Größe von 541 Terrabyte. [51]

[48] Vgl. Bertin-Mahieux, T. (2014)
[49] Vgl. Wikimedia Foundation (o.J.)
[50] Vgl. Amazon Web Services (2015b)
[51] Vgl. Amazon Web Services (2013)

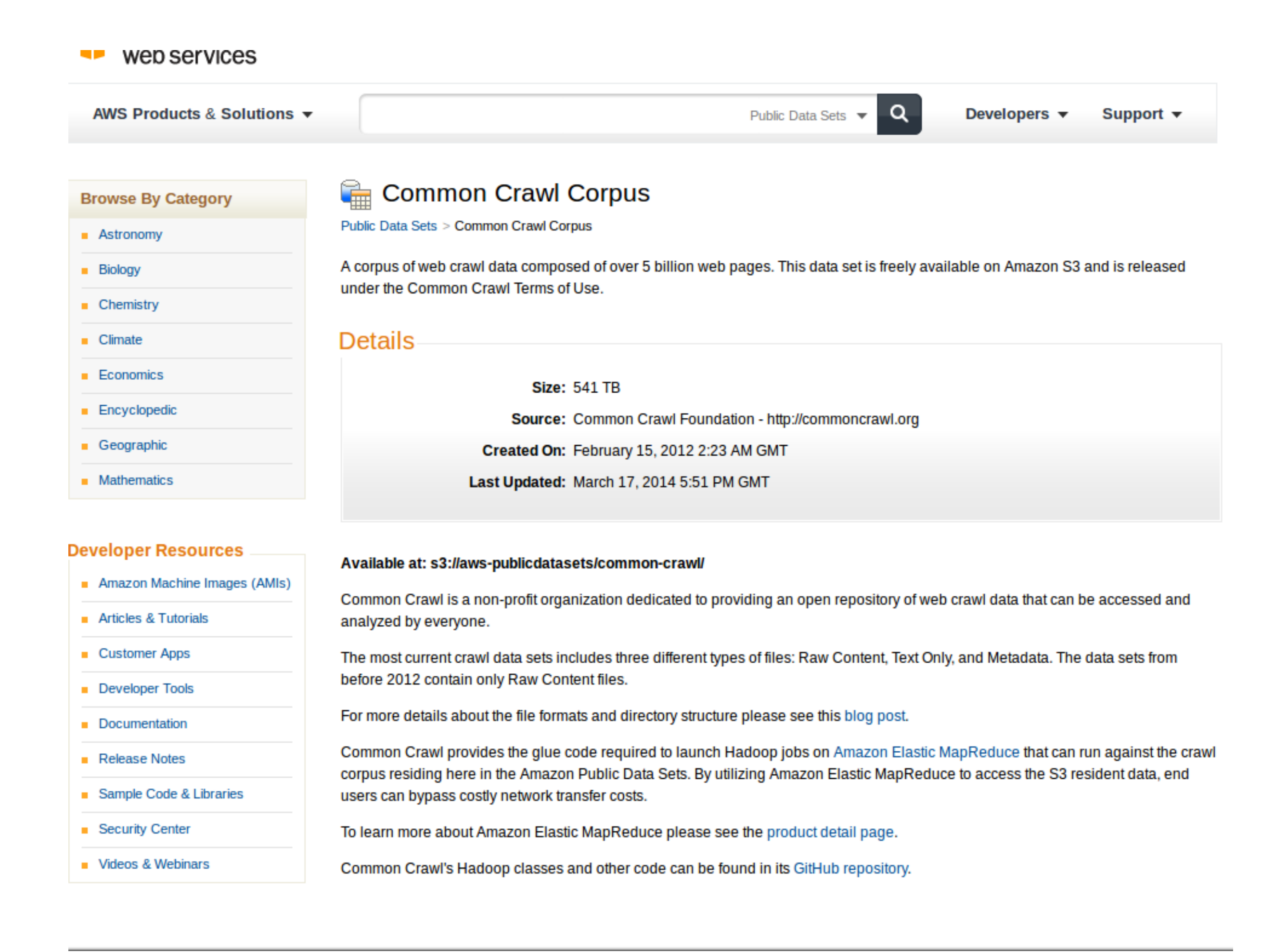


Abb.4: Common Crawl Datenset auf AWS Homepage

Die Verwendung der Common Crawl Database konnte letztendlich innerhalb Cloudant jedoch nicht durchgeführt werden, da die Migration von Amazon AWS mit den vorhandenen Ressourcen – vor allem aufgrund monetärer Gesichtspunkte – nicht umsetzbar war. Die Migration der Daten von AWS nach Cloudant wäre zwar möglich gewesen, hätte aber immense Kosten verursacht.

Dies machte es für unsere Tests unmöglich, die Daten in Cloudant zu importieren und auszuwerten. Hier gilt allerdings zu erwähnen, dass ein solches Szenario für die Praxis zudem untypisch ist, da in der Regel keine großen bestehenden Datenmengen verwendet werden, sondern dynamisch wachsende Datensets entstehen, weshalb im Falle von Cloudant auf ein Testszenario der Daten-Aggregierung zurückgegriffen wurde.

Aggregation von Daten in Cloudant

Da die Replikation eines Big-Data-Datensatzes innerhalb Cloudants wie bereits beschrieben mit den vorhandenen Ressourcen nicht praktikabel war, wurde die Not zur Tugend gemacht: Um neben der Größe des Datensatzes auch das Datenwachstum sowie die Menge an komplexen Abfragen zu betrachten, wird innerhalb Cloudants deshalb ein Datensatz aggregiert.

Mithilfe des IBM-Produktes „Bluemix“, werden alle Twitter-Tweets mit dem Trending-Hashtag „JeSuisCharlie“ in die Datenbank im JSON-Format abgespeichert. Die genaue Erläuterung der technischen Umsetzung erfolgt in Kapitel 3.3.

Betrachtung des Dataset-Frameworks

Auch wenn eine große Anzahl von Datensätzen öffentlich verfügbar ist, so ist es durchaus eine komplexe Aufgabenstellung, Datensätze zu finden, die auch als “Big-Data”-Datensatz gelten – also die Leistungsfähigkeit herkömmlicher relationaler Datenbanksysteme deutlich überschreiten. Dabei gibt es jedoch keine exakte Definition, ab wann ein Datensatz als ein “Big-Data”-Datensatz zu bewerten ist.

Das Framework und deren Anforderungen sind folglich ein erster Schritt, um in Zukunft die Unterscheidung zwischen herkömmlichen und neuen Datenbanksystemen voranzutreiben. Aktuell werden sehr häufig vergleichsweise kleine Datensätze mit nicht-relationalen Datenbanksystemen ausgewertet, die jedoch auch nach Aussage deren Entwickler besser mit herkömmlichen relationalen Datenbanklösungen ausgewertet werden sollten.

Dabei gilt vor allem, die effiziente Nutzung der Big-Data-Technologie nicht aus den Augen zu verlieren. Wo oft von “Big Data” geredet wird, wird eigentlich “Small Data” gemeint, denn die Datenmengen sollten nicht nur extrem groß sein, sondern können auch unstrukturiert und in großer Geschwindigkeit auftreten. Das Framework hilft folglich bei der Prüfung, ob ein Big-Data-Ansatz überhaupt sinnvoll ist. In der Regel treten große und unstrukturierte Datenmengen nämlich nicht als fertiger Datenbanksatz auf, sondern diese müssen zumeist zunächst aggregiert werden. [52]

3.2 Testszenario 1 – Analyse der Common Crawl DB mit Amazon EMR

Im ersten Testszenario wird mit Hilfe von Amazon Elastic Map Reduce die Anzahl der Wörter innerhalb des bereits beschriebenen Common Crawl Datensatzes gezählt. Das Szenario orientiert sich an der Vorgehensbeschreibung aus dem dazugehörigen Common Crawl Blog.[53]

Zunächst muss in Amazon AWS ein Account eingerichtet werden. Im Anschluss wird lokal auf dem Client für die Hadoop Programmierung die IDE “Eclipse” installiert, sowie der Elastic Map Reduce Ruby Client.

[52] Vgl. Gloster, F. (2014)
[53] Vgl. Salevan, S. (2011)

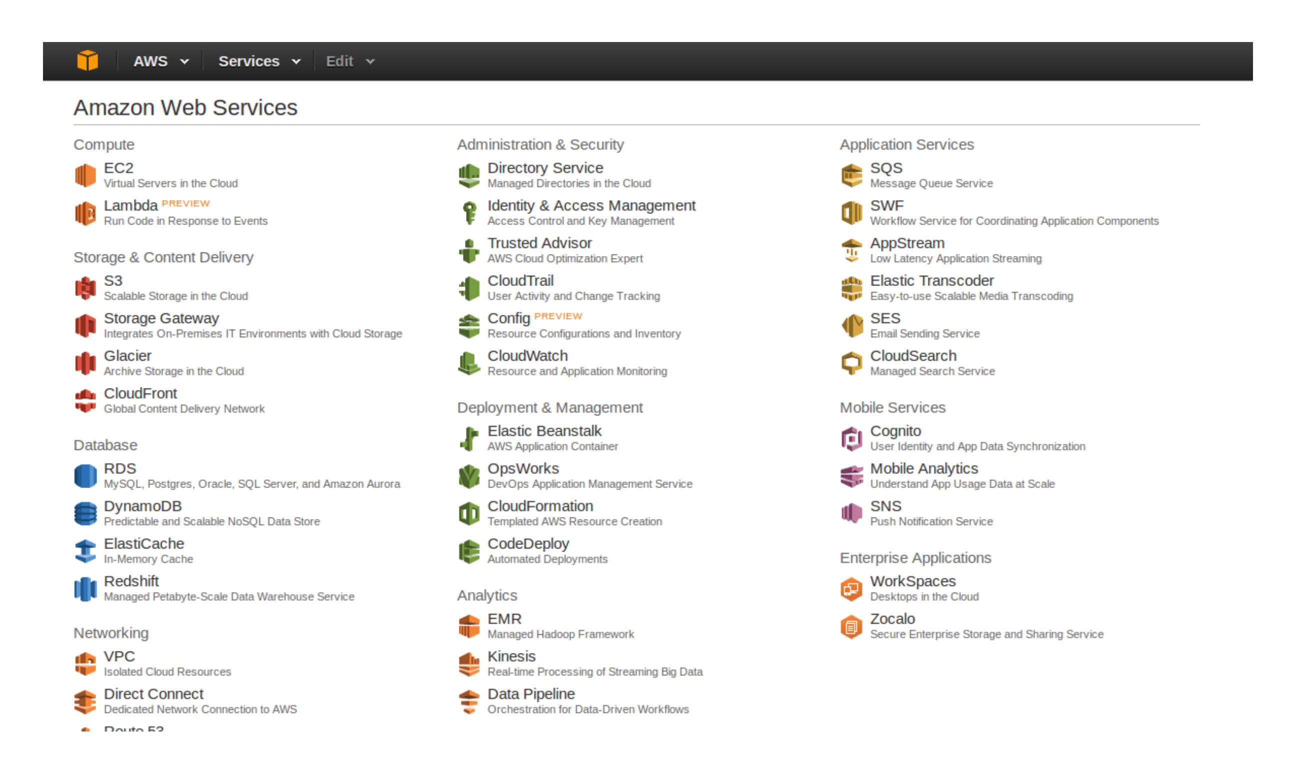


Abb.5: AWS Dashboard

Im Anschluss wird eine Amazon Elastic Compute Cloud (EC2) Instanz aufgesetzt und zusätzlich ein sogenannter „Bucket“ im Amazon Simple Storage Service erstellt – ein Bucket ist ein Ordner, in dem Dateien auf dem Server gespeichert werden können.

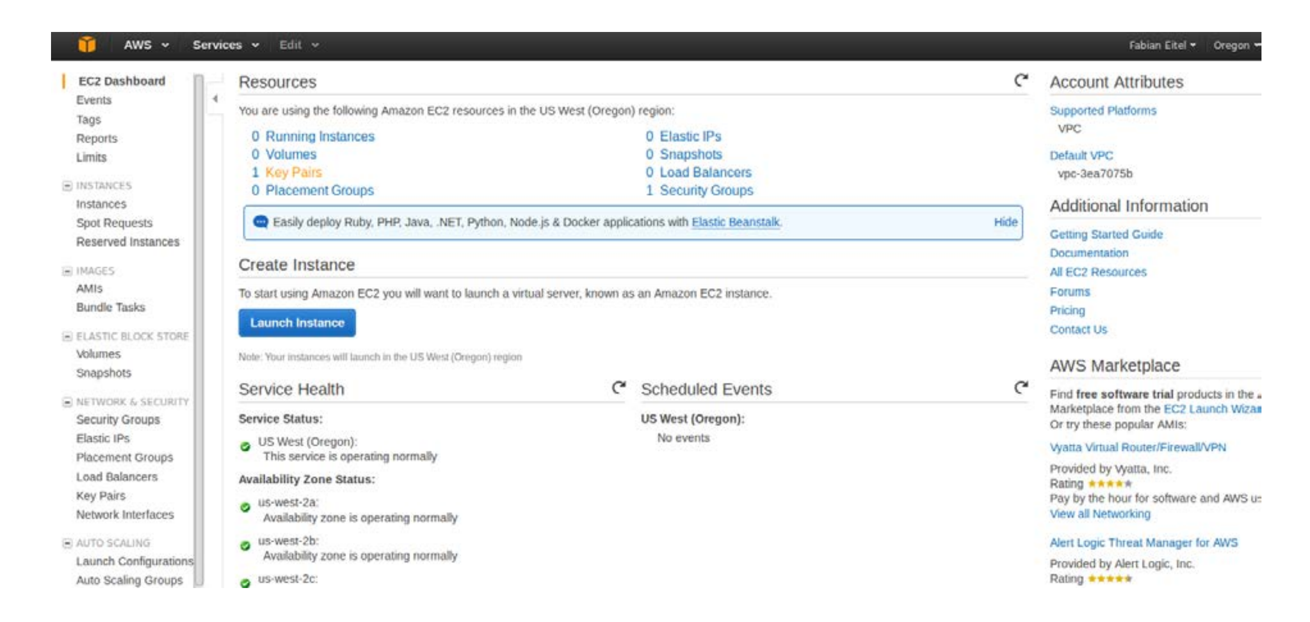


Abb.6: EC2 Instanzen aktiviert

Im nächsten Schritt muss der Elastic Map Reduce Ruby Client konfiguriert werden. Der Client benötigt die Zugangsdaten in Form einer JSON-Datei. Die besagte Datei *credentials.json* wird in das gleiche Verzeichnis wie der Ruby Client gespeichert.

Abb.7: Inhalt der credentials.json Datei

Mit dem Terminal (in unserem Beispiel wird die Linux Distribution „Mint“ verwendet), wird nun der CommonCrawl „Hello World“ Beispielcode ausgeführt, der es ermöglicht, die Anzahl der Wörter innerhalb des Datensatzes auszugeben. Dieser Beispielcode ist auf Github vom Nutzer „ssalevan“ erstellt worden und ist frei im Internet abrufbar.

Abb.8: Download des Codes von Github

Nun wird Eclipse gestartet, um ein neues Projekt aus den heruntergeladenen Programmcode-Dateien in Form einer .JAR Datei zu kompilieren:

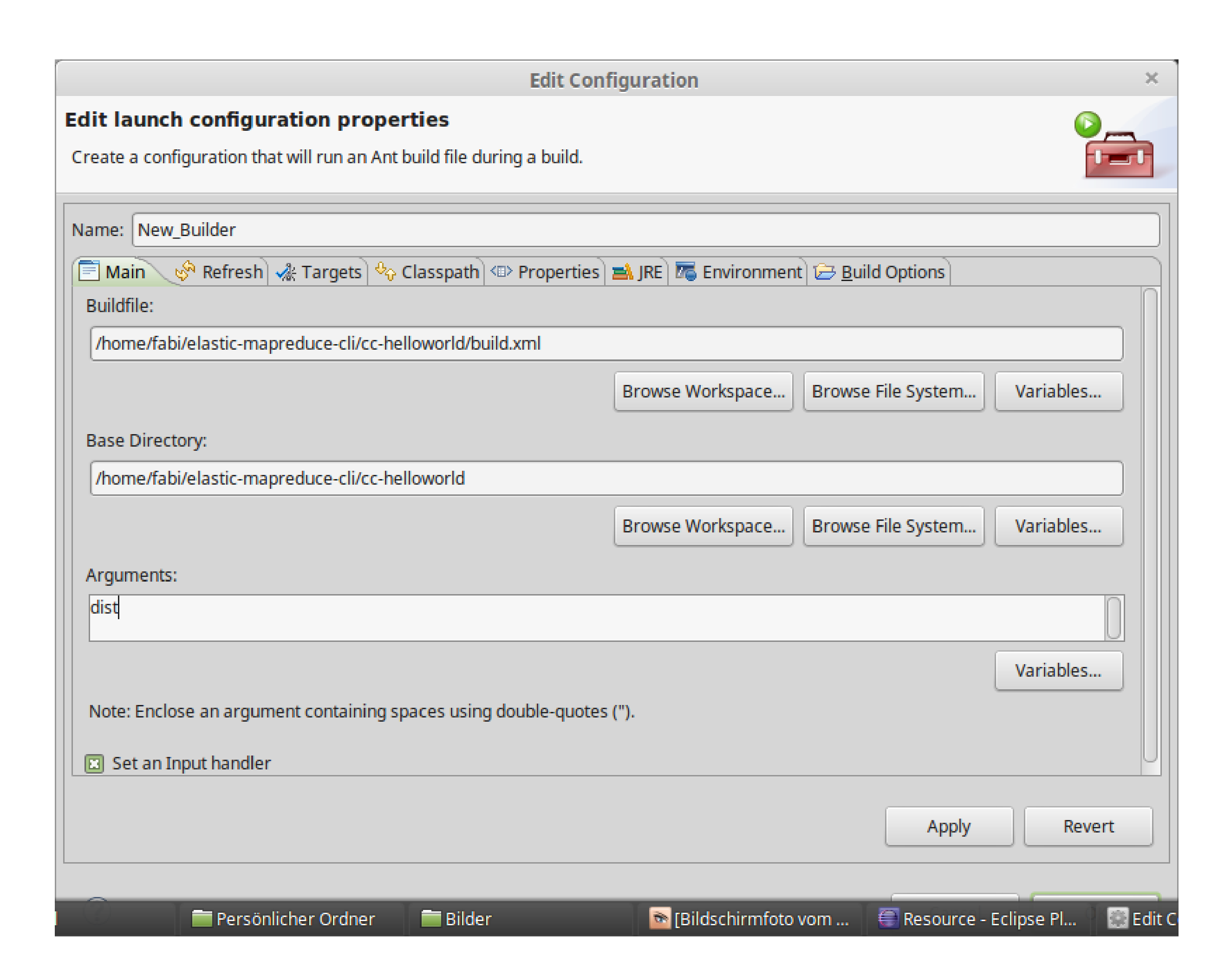


Abb.9: Definition des Builders in Eclipse

Die mit Eclipse kompilierte .JAR Datei wird dann auf den Amazon S3 Bucket hochgeladen. Danach kann ein Elastic Map Reduce Job basierend auf der .JAR Datei gestartet werden.

```
Terminal
     testlauf
j-N38HJVVFVIBY      FAILED
     testlauf
fabi@fabi-mint ~/elastic-mapreduce-cli $ ./elastic-mapreduce --jobflow j-Q77H720
A7GQR \
>  --jar s3n://nosql-dhbw/HelloWorld.jar \
>  --arg org.commoncrawl.tutorial.HelloWorld \
>  --arg AKIAJEDKFMM3NANBYBWQ \
>  --arg 2YEGZoaBJXNV3Bf9dNjBhH4WTRvazfiBBz9etclr \
>  --arg common-crawl/crawl-002/2010/01/07/18/1262876244253_18.arc.gz \
>  --arg s3n://nosql-dhbw/helloworld-out
Added jobflow steps
fabi@fabi-mint ~/elastic-mapreduce-cli $ ./elastic-mapreduce --list
j-Q77H720A7GQR      BOOTSTRAPPING  ec2-54-208-2-244.compute-1.amazonaws.com
     testlauf
   PENDING        Example Jar Step
```

Abb.10: Erstellung des MapReduce Jobs und Auflistung

Der oben stehende Kommandozeilen-Code führt folgende Dinge aus:

- Führe die main() Methode in der HelloWorld Klasse aus
- Login in den Amazon S3 mit dem AWS Zugangscode.
- Zähle alle Wörter die am 07.01.2010 um 18:00 Uhr vom Crawler heruntergeladen worden sind. Die Common Crawl Crawling-Daten werden im ARC-Format komprimiert gespeichert – deswegen die Dateiendung arc.gz.
- Ausgabe der Ergebnisse als eine CSV-Datei in den Amazon S3 Bucket (das Verzeichnis lautet helloworld-out).

Auf diesem Screenshot ist ersichtlich, wie drei EC2 Instanzen den Elastic Map Reduce Job seit 44 Minuten ausführen.

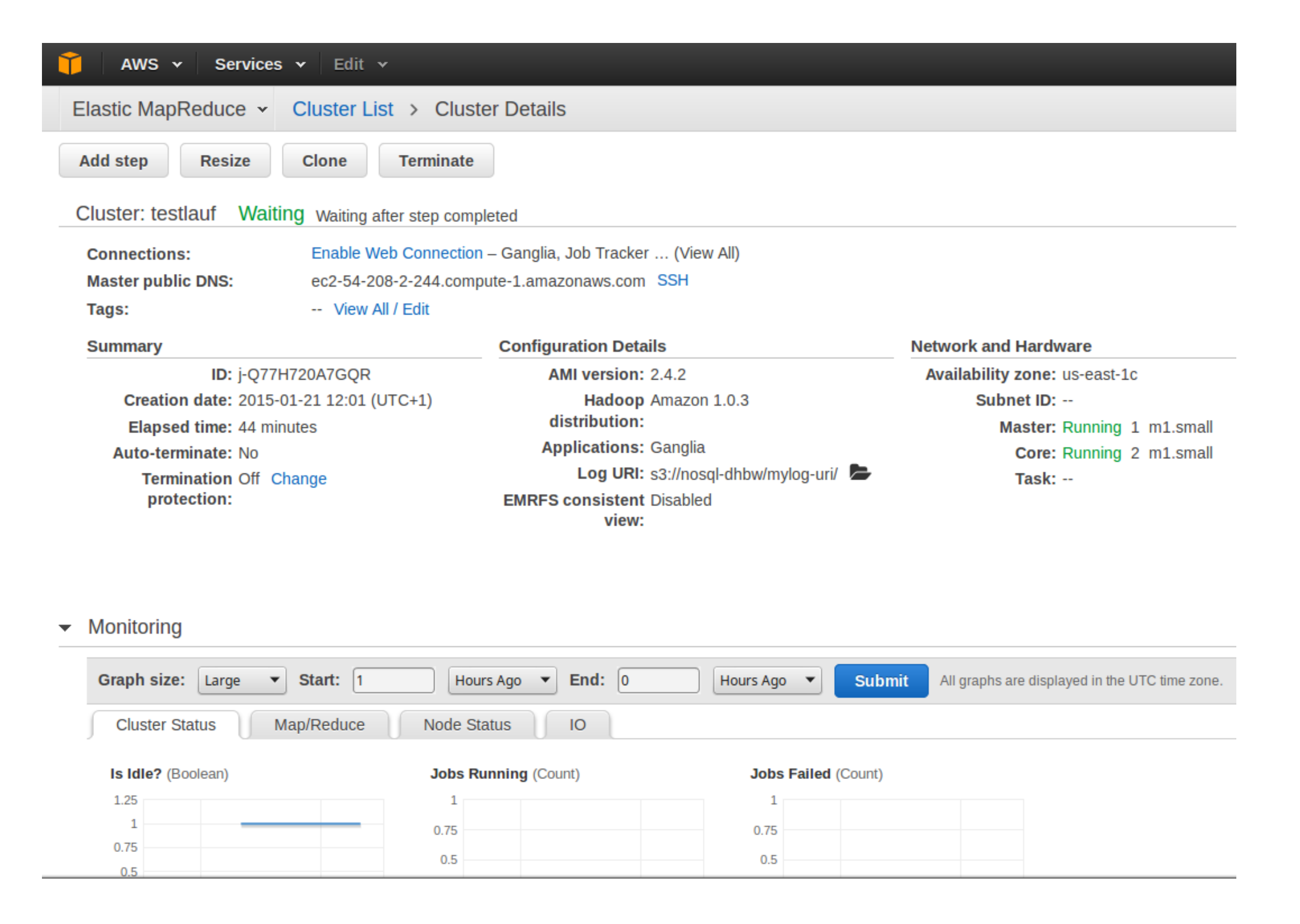


Abb.11: Cluster Übersicht

Je nach Größe des Datenabschnittes und der Anzahl der Serverinstanzen, variiert die Analysedauer der Daten. Am Ende wird die CSV-Datei in den S3 Bucket gespeichert und kann vom Benutzer heruntergeladen werden. Die Ergebnis Datei zeigt nun jedes aufgetretene Wort und die absolute Häufigkeit der Nennung. Groß- und Kleinschreibung wird hierbei nicht beachtet. Per Volltextsuche können nun einzelne Begriffe analysiert werden.

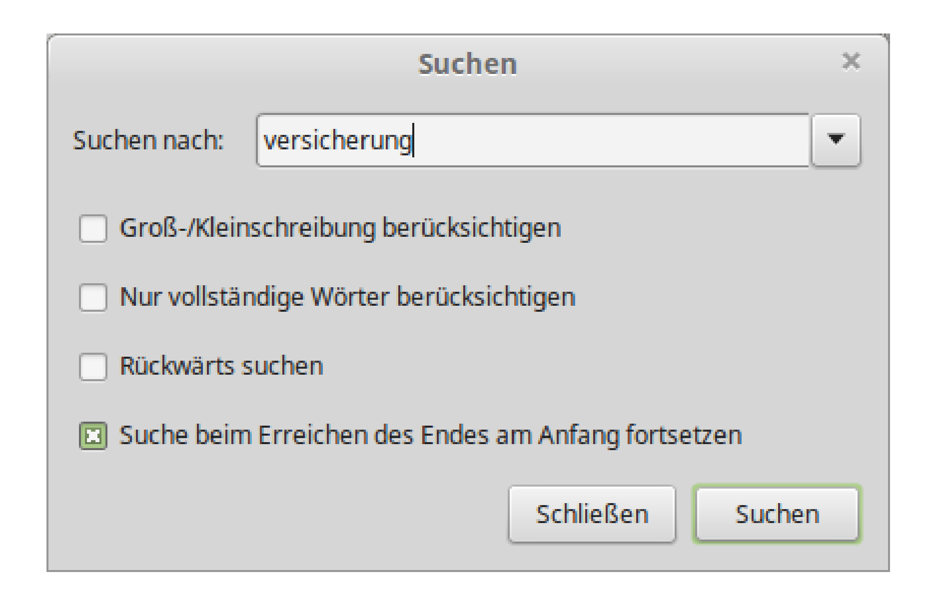


Abb.12: Volltextsuche in der Ergebnis-Datei

Die Gesamtabfrage des spezifischen Crawls vom 07.01.2010 konnte auf dem verwendeten EMR Clusters, bestehend aus 1 Master und 2 Slaves der EC2 Small Instanzen, innerhalb von einer Viertelstunde ausgeführt werden. Den größten Teil der Zeit beansprucht hierbei das Starten des Clusters und nicht die Abfrage selbst. In einem weiteren Versuch wurde der gesamte Datensatz aus dem 2010 Crawl untersucht der ungefähr 100TB beträgt und 3,8 Milliarden Seiten umfasst. Innerhalb von einer Stunde konnten ca. 2 GB der Daten verarbeitet werden, rechnet man dies auf den gesamten Datenbestand hoch, würde die Berechnung 50.000 Stunden benötigen. Alternativ könnte man das 50.000-fache an Serverleistung oder schnellere Instanzen verwenden. Aufgrund der stündlichen Preisstruktur von AWS würden die ersten beiden Möglichkeiten die gleichen Kosten verursachen. Die Kosten für die Laufzeit der Tests betrugen 0,90$ von denen die EC2 Instanzen 0,40$ ausmachen (siehe Abb.13).

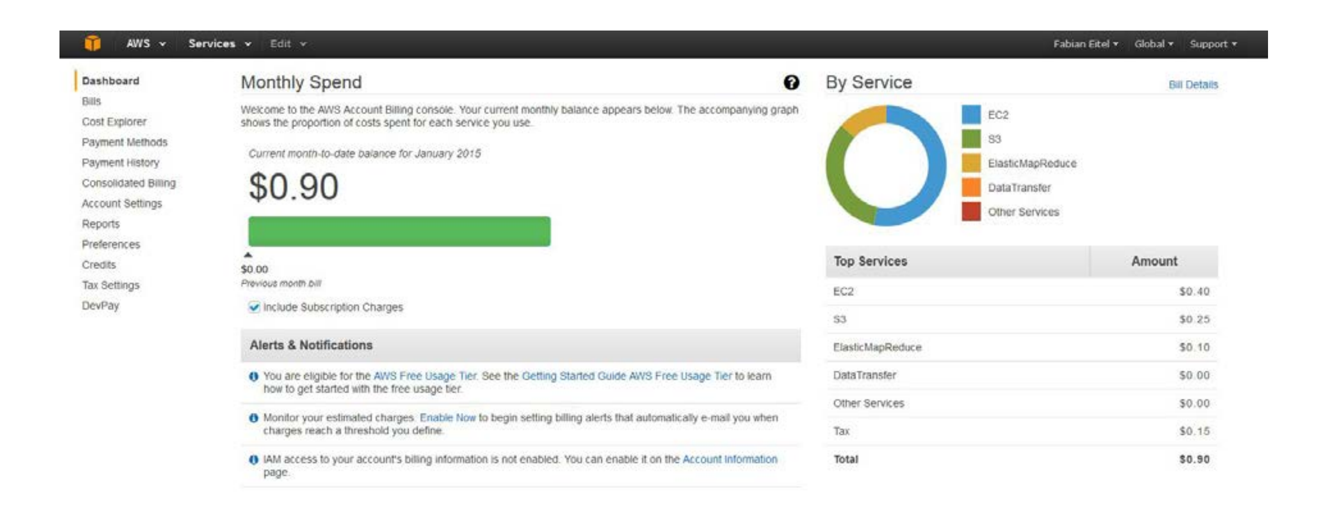


Abb. 13: Ausgabenübersicht

Zusammenfassend betrachtet ist die Komplexität einer Amazon EMR Anwendung hoch. Vom Anwender wird ein detailliertes Wissen über Hadoop, MapReduce und die Amazon Web Services Umgebung vorausgesetzt. Viele Aktivitäten lassen sich ausschließlich über die Konsole lösen und fordern die Installation von spezifischen Programmen und Sprachen. So unterstützt AWS z.B. lediglich die Java Version 6 und nicht die aktuellere Version 7. Unternehmen benötigen daher für die Benutzung von Amazon EMR entweder hoch qualifizierte Spezialisten für Big Data Lösungen oder müssen auf externe Dienstleister zurückgreifen.

3.3 Testszenario 2 – Regionale Analyse von Tweets mit Cloudant

Die Registrierung erfolgt bei Cloudant über ein Webformular und ist zunächst kostenlos, auch die Eingabe einer Kreditkartennummer wird, anders als bei Amazon, nicht direkt benötigt. An kostenlosen Kapazitäten bietet Cloudant im Multi-Tenant Modus, bei dem sich mehrere Kunden einen Datenbankserver teilen, 50$ pro Monat an (Stand 15.01.2015).

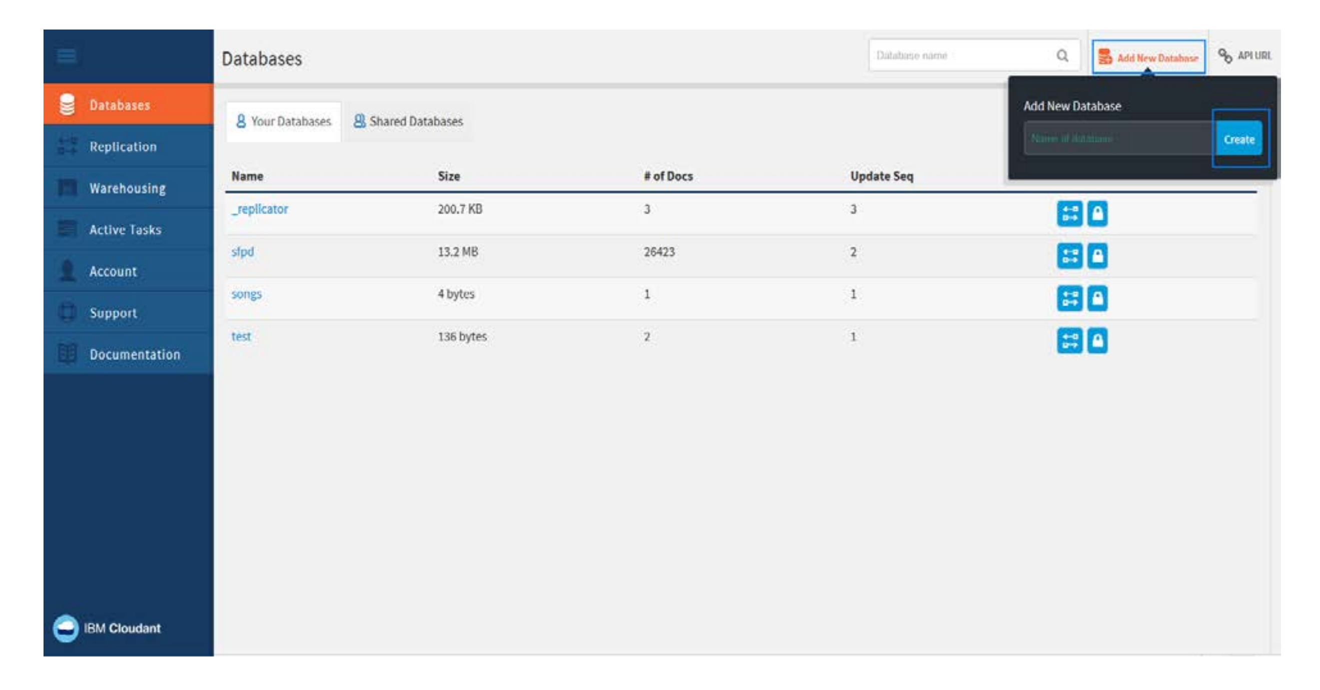


Abb.14: Erstellung einer Cloudant Datenbank

Nach dem Erstellen des Cloudant Accounts wird im nächsten Schritt eine erste Datenbank angelegt. Dies benötigt lediglich zwei Klicks sowie die Eingabe des Namens (siehe Abb. 14) und geschieht dann innerhalb von wenigen Sekunden.

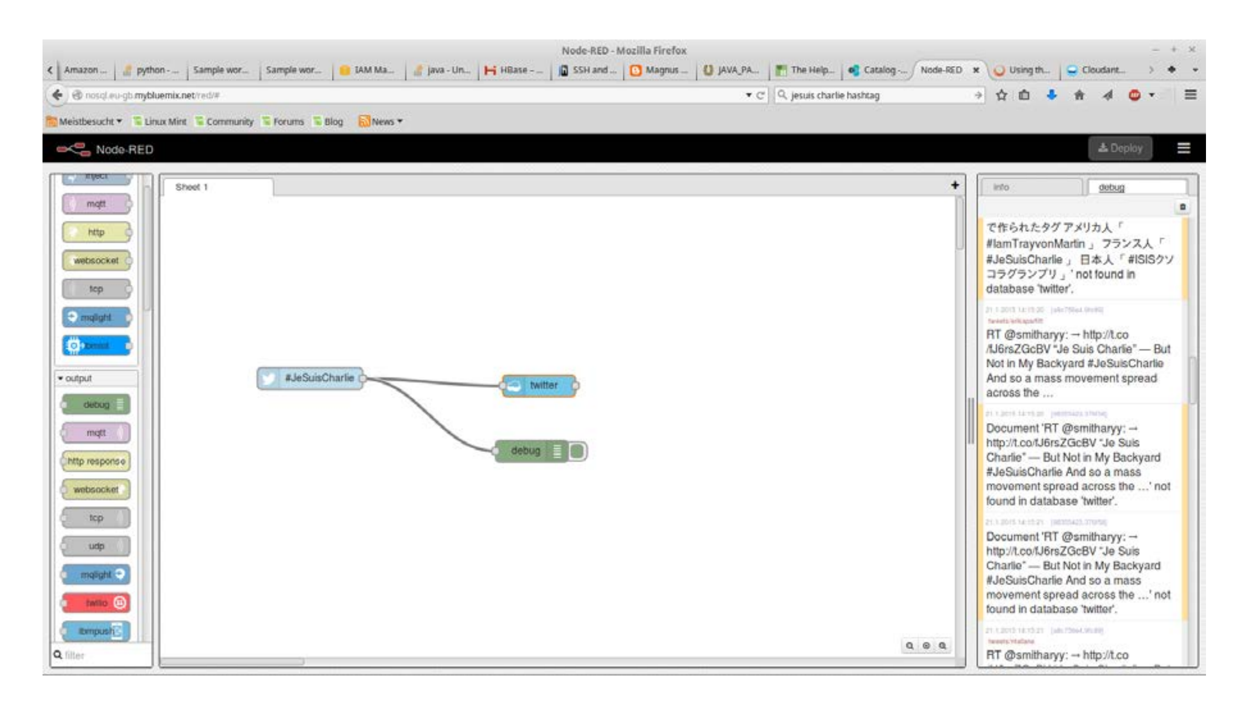


Abb. 15: Verbindung von Cloudant und Twitter API in Cloudant

Zur Erstellung der Testdaten wird eine Internet-of-Things Broilerplate in der PaaS IBM BlueMix angelegt. Mithilfe des beinhalteten Tools NodeRed verbindet BlueMix die Cloudant Datenbank über eine grafische Oberfläche mit der Twitter API (siehe Abb. 15). Das hiermit erstellte Tool kann nun direkt auf entsprechender Infrastruktur installiert und getestet werden. Innerhalb von drei Tagen wurden ca. 300 MB in knapp 40.000 Dateien, respektive Posts, gesammelt.

Auf die selbst erhobenen Daten in der Datenbank von Cloudant sollen nun sinnvolle Abfragen geschaltet werden. Dazu ist es vorteilhaft, sich die Struktur eines repräsentativen JSON-Dokuments des Twitter Datenbestandes anzuschauen.

 https://bluemchen.cloudant.com/twitter/f5420e9d963370ff2ccc174773ff6f0e

Abb.16: URI eines JSON-Dokuments

Jedes Dokument in der Datenbank ist mit einer eindeutigen ID und darauf aufbauend einem Uniform Resource Identifier (URI) ausgewiesen. Wie in Abbildung 16 dargestellt, kann hiermit das JSON-Dokument im Internetbrowser angezeigt werden. Die URI setzt sich in diesem Fall aus der Adresse des Cloudant Accounts, dem Datenbanknamen und der ID zusammen. Ein solches JSON-Dokument aus dem Twitter Datenbestand besteht aus 386 Zeilen und verdeutlicht, dass hier viele Key-Value-Paare enthalten sind, die abgefragt werden können.

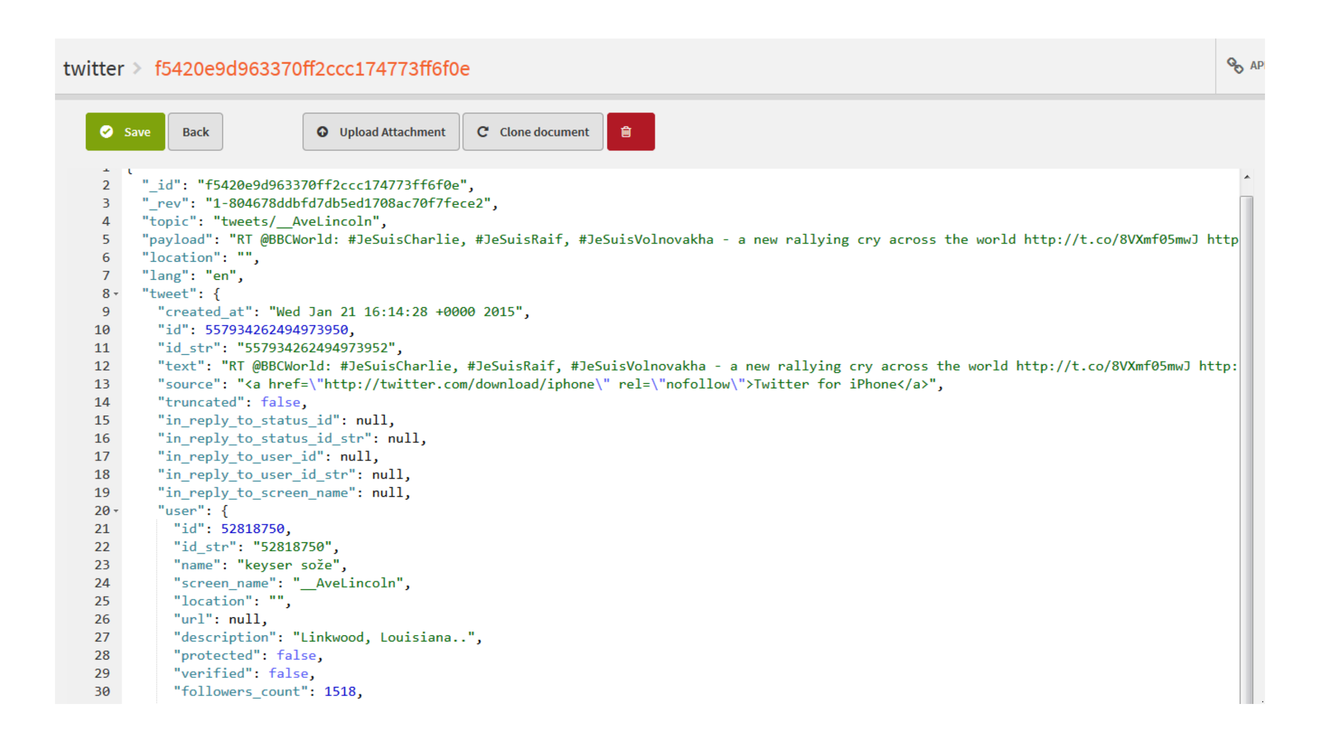


Abb.17: JSON-Dokument der Twitterdatenbank in Cloudant

Natürlich kann man sich das JSON-Dokument auch in Cloudant anzeigen lassen, wie in Abbildung 17 gezeigt. Diese formatierte Ansicht ist deutlich übersichtlicher als die nicht formatierte Anzeige im Internetbrowser und ist daher für die vorliegende Untersuchung die bevorzugte Methode.

In den JSON-Dokumenten werden unter anderem der Twitter Username, die Sprache und die Zeitzone angegeben. Insgesamt sollten drei Testfälle durchgeführt werden.

Beim ersten Testfall soll für jede im Twitter Datenbestand vorhandene Zeitzone die absolute Häufigkeit ermittelt werden. Dementsprechend muss die Abfrage alle Tweets aus den jeweiligen Zeitzonen aggregieren und die Gesamtzahl ausgeben. Damit kann die geografische Verteilung der Tweets beurteilen werden. Für die Abfrage muss ein View in Cloudant erstellt werden. Dieser View wird auf jedes JSON-Dokument im Twitter Datenbestand angewendet. Um sich die Ergebnismenge des Views anzeigen zu lassen, muss der View abgefragt werden. Der Nutzer muss den View also nicht selber ausführen. Prinzipiell kann der View in verschiedenen Programmiersprachen umgesetzt werden. Aufgrund der bereits bestehenden Kenntnisse und der einfachen Umsetzung von JavaScript in Cloudant, wurde dieses für die Umsetzung der Views verwendet.

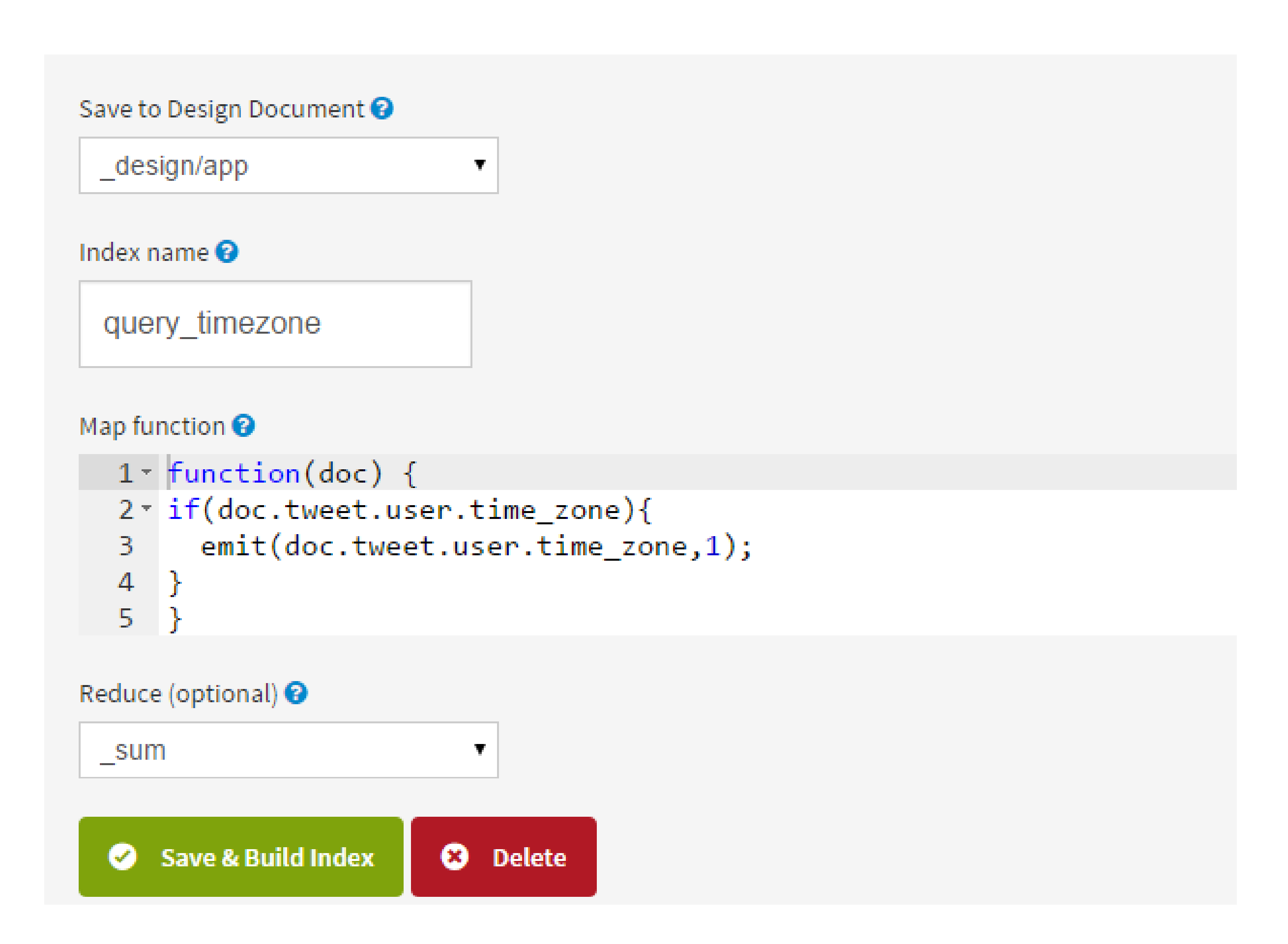


Abb.18: MapReduce-Funktion zur Durchführung des Testfalls

Abbildung 18 zeigt den Code für den in diesem Testfall verwendeten View. Der View Index wird in einem Design Dokument mit dem Namen „_design/app“ gespeichert. An dieser Stelle werden alle View Indize gespeichert, die für diese Untersuchung verwendet werden. Der Name des Indexes ist „query_timezone“. Hierbei handelt es sich um eine MapReduce-Funktion. In dieser Ausführung prüft die Funktion, ob ein Wert für die Zeitzone vorhanden ist. Trifft das zu, dann wird als Key die jeweilige Zeitzone ausgegeben und als Value die 1 gesetzt. Über den Reduce-Teil der Funktion werden die Werte zusammengezählt, wodurch die absolute Häufigkeit der Tweets in einer Zeitzone ausgegeben wird. Speichert man den View ab, wird er automatisch auf jedes Dokument in der Twitter Datenbank angewendet.

https://bluemchen.cloudant.com/twitter/_design/app/_view/query_timezone?group=true


```
{"rows":[
{"key":"Abu Dhabi","value":57},
{"key":"Adelaide","value":7},
{"key":"Alaska","value":20},
{"key":"Almaty","value":3},
{"key":"America/Los_Angeles","value":1},
{"key":"America/Mexico_City","value":1},
{"key":"America/New_York","value":6},
{"key":"Amsterdam","value":1157},
{"key":"Arizona","value":84},
{"key":"Asia/Calcutta","value":2},
{"key":"Athens","value":1468},
```


Abb.19: Absolute Häufigkeit der Tweets in einer Zeitzone

In Abbildung 19 wird gezeigt, wie man anschließend den View mittels eines HTTP GET-Befehls im Internet Browser abrufen und sich so die Ergebnismenge ausgeben lassen kann – ebenfalls im JSON-Format.

Der zweite Testfall soll die Gesamtzahl aller Tweets in französischer Sprache im Twitter Datenbestand anzeigen. Hierfür wird ebenfalls eine MapReduce-Funktion mit JavaScript erstellt.

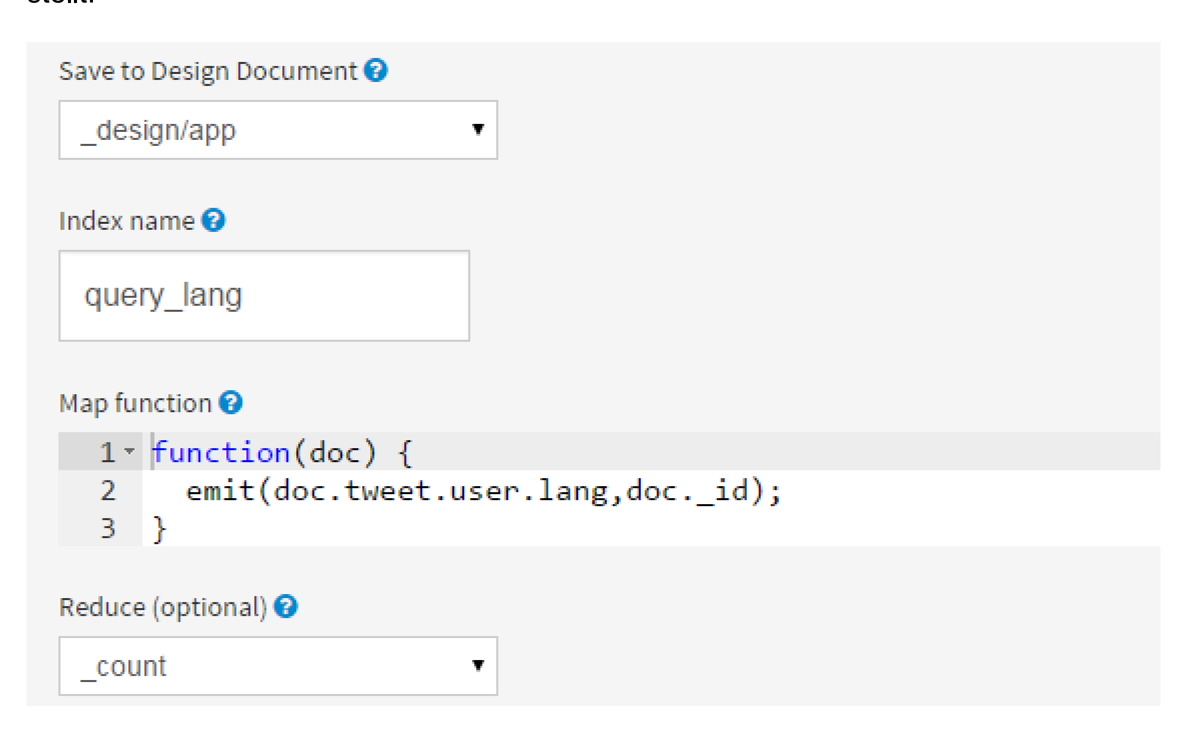


Abb.20: MapReduce-Funktion zur Anzeige der Sprache der Tweets

Abbildung 20 zeigt die Abwandlung der Funktion des vorherigen Testfalls. In diesem Fall wird als Key die Sprache des Tweets verwendet und als Value die ID des JSON-Dokuments zurückgegeben. Anschließend werden im Reduce-Schritt die Anzahl der Zeilen mit der gleichen Sprache gezählt. Der Name dieses View Indexes ist „query_lang“. Mit einem HTTP GET-Befehl kann hier ebenfalls die Ergebnismenge des Views ausgegeben werden.

https://bluemchen.cloudant.com/twitter/_design/app/_view/query_lang?group=true&key="fr"

{"rows":[
{"key":"fr","value":15267}
]}

Abb.21: Ausgabe der Gesamtzahl an Tweets in Französisch

Wie in Abbildung 21 dargestellt, wird hier die Ergebnismenge des Views nach dem Key „fr" gefiltert. Somit bekommen wir ausschließlich den Wert für diesen Key angezeigt. Das zeigt, wie einfach es ist die Ausgabe der Views komfortabel mittels Befehl der klassischen HTTP API auf die eigenen Bedürfnisse anzupassen.

Diese Testfälle zeigen sehr schön, wie einfach die Implementierung von Views und die anschließende Ausgabe der Ergebnismenge mittels HTTP-Befehlen ist. Das ist vor allem für die Verwendung mit Webapplikationen hilfreich. Diese Eigenschaft lässt sich vor allem darauf zurückführen, dass Cloudant auf CouchDB basiert. Dieses hat als Ziel und Eigenschaft die einfache Verwendung mit Webtechnologien, folglich trifft das auch auf Cloudant zu. Auf wirtschaftlicher Seite ist wichtig zu wissen, wie viel die Nutzung von Cloudant kostet.

Data volume in GBs / month	**$1.00 per GB / month**
"Heavy" API requests * PUTs, POSTs, DELETEs	**$0.015 per 100**
"Light" API requests * GETs, HEADs	**$0.015 per 500**

Abb.22: Preisliste für die Nutzung von Cloudant (Stand 15.01.2015)[54]

Abbildung 22 zeigt, wie viel „leichte" und „schwere" API Anfragen kosten. Darüber hinaus zahlt man für jeden GB Datenbestand pro Monat. Hier einschränkend muss erwähnt werden, dass die Nutzung kostenlos ist, solange weniger als $50 pro Monat anfallen. Innerhalb dieser Untersuchung wurden sowohl „leichte" als auch „schwere" API Anfragen verwendet, die jedoch unter den angegebenen Werten von 100 und 500 blieben. Da ein Teil der Autoren für die IBM Deutschland GmbH tätig ist, konnte Cloudant kostenlos genutzt werden. Die im System angezeigten Kosten und Nutzungswerte sind aggregiert von allen IBM-Mitarbeitern, die ebenfalls Cloudant kostenlos auf dem gleichen Cluster verwenden. Für die vorliegende wissenschaftliche Arbeit ist diese Analyse der Kosten also nicht repräsentativ. Das Volumen der

[54] Enthalten in: Cloudant (2014)

Views für den Twitterdatenbestand betrug 6 MB bei einer Testdatenbankgröße von ungefähr 300 MB. Für NoSQL-Datenbanksysteme wie Cloudant sind das geringe Größen, wodurch die Nutzung auch ohne die Unterstützung der IBM Deutschland GmbH kostengünstig ausgefallen wäre.

Zum Zeitpunkt der Untersuchung bestand der Twitter Testdatenbestand aus fast 40.000 JSON-Dokumenten – Tendenz stark steigend. Das ist eine natürliche Entwicklung, die bei NoSQL-Datenbanken beziehungsweise bei Big Data häufig auftritt. Die Datenbank erhält nicht sofort alle Daten beim Aufsetzen, sondern wächst organisch. Für den ersten Testfall war der HTTP GET-Abruf der Ergebnismenge 5 kB groß und hat 465ms benötigt. Beim zweiten Testfall, der deutlich spezifischer war, war der Abruf 42 Byte groß und hat 155ms benötigt.

4 Fazit und Ausblick: Anwendungsfälle von NoSQL-DB

4.1 Fazit

Die vorliegende Arbeit unterstreicht die wachsende Bedeutung von NoSQL-Datenbankensystemen – insbesondere die Bereitstellung dieser Systeme über die Cloud ist sehr komfortabel. Die Untersuchung hat gezeigt, dass die Anbieter von DBaaS grundsätzlich einen hohen Kundenfokus aufweisen. Die Anmeldung zu den Diensten kann schnell durchgeführt werden. Des Weiteren hat die vorliegende Arbeit gezeigt, dass die Kosten für die vorgestellten Testszenarien geringfügig sind.

Im Vergleich der Testszenarien von Amazon und Cloudant fällt auf, dass das Aufsetzen einer Datenbank bei Letzterem deutlich einfacher und komfortabler ausfällt. Mit wenigen Schritten können Datenbanken repliziert und erste Abfragen geschaltet werden. Bei Amazon ist der Einstieg erschwert, was vor allem daran liegt, dass zahlreiche Skripte installiert werden müssen, bevor mit dem Aufsetzen der Datenbank und den ersten Abfragen begonnen werden kann, beziehungsweise sogar Code programmiert werden muss. Letztlich hat die Ausarbeitung gezeigt, dass vor allem bei Datenbanken aus der Cloud die Benutzerfreundlichkeit und eine transparente Preisstruktur für den Nutzer wichtig sind. Sowohl bei Amazon EMR als auch bei Cloudant werden die Preise anhand der Datenbankgröße und den jeweiligen Abfragen bemessen, wobei die Kosten erst für große Datensätze relevant werden.

Abschließend lässt sich festhalten, dass die NoSQL-Datenbanksysteme aus der Cloud insbesondere dann sinnvoll sind, wenn man einen hohen Wert auf Flexibilität legt. Die Testdaten werden einfach in die Datenbank geladen, ohne vorher deren Struktur definieren zu müssen. Wie in den Testfällen festgestellt, ist das vor allem bei Daten aus sozialen Netzwerken enorm wichtig. Im Vorfeld kann nicht eingeschätzt werden welche Struktur die Daten haben werden – dementsprechend problematisch wäre hier die Verwendung von relationalen Datenbanksystemen.

4.2 Ausblick

Die Untersuchung hat gezeigt, dass bereits jetzt NoSQL-Datenbanksysteme in der Cloud vorhanden sind, die einen schnellen und kostengünstigen Einstieg in die Materie erlauben. Der exponentielle Anstieg des analysierbaren Datenvolumens stellt eine neuartige Möglichkeit dar, interessante wirtschaftliche Informationen gewinnen und daraus letztendlich eine bessere Entscheidungsgrundlage für unternehmerische Entscheidungen erhalten zu können.

Insbesondere das Web und damit einhergehend soziale Netzwerke spielen hier eine große Rolle und werden in Zukunft noch bedeutender werden. Über die Analyse von sozialen Netzwerken können sehr genau die Empfindungen und Bedürfnisse von Gruppen erörtert werden. Daraus ergeben sich vor allem fürs Marketing und für den Vertrieb große Chancen. Mit Daten aus sozialen Netzwerken und dem Nutzerverhalten können in Zukunft Streuverluste beim Marketing verringert und die Kundengruppe einfacher und genauer erreicht werden. Das ermöglicht ebenfalls Produkt- sowie Service-Angebote konkreter auf die jeweiligen Kundensegmente abzustimmen.[55]

[55] Vgl. Weber, M. (2012), S. 34

Anhang

Quellenverzeichnisse

Literaturverzeichnis

Baun, C. / Kunze, M./Nimis, J./ Tai, S. (2011): Cloud computing web-based dynamic IT services,Springer: Berlin Heidelberg

Frampton, M. (2015): Big data made easy, Apress: New York City, S. 1-10

IBM (2014): Technical Overview: Anatomy of the CloudantDBaaS, IBM Corporation: Somers

Kuznetsov, S. D. / Poskonin, A. V. (2014): NoSQL data management systems, in: Programming and Computer Software, 2014, Vol. 40, Nr. 6, S. 323 - 332

Lenk, A. / Klems, M. /Nimis, J./ Tai, S. (2009): What's Inside the Cloud?: An Architectural Map of the Cloud Landscape, Hewlett-Packard Laboratories: Palo Alto

National Institute of Standards and Technology (2011): The NIST Definition of Cloud-Computing, in: Computer Security, 2011, S. 1-3

Redmond E./Wilson J.(2012): Sieben Wochen, sieben Datenbanken: Moderne Datenbanken und die NoSQL-Bewegung, O'Reilly Verlag: Köln, S.100-120

Wadkar, S./Siddalingaiah, M. (2014): Pro Apache Hadoop, Zweite Auflage, Apress: New York City, S. 1-20

Weber, M. (2012): Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte (Hrsg.: BITKOM), BITKOM: Berlin-Mitte

Verzeichnis der Internet- und Intranet-Quellen

Amazon Web Services (2013): Common Crawl Corpus, https://aws.amazon.com/datasets/41740, Abruf: 20.01.2015

Amazon Web Services (2015a): Amazon EMR, http://aws.amazon.com/elasticmapreduce/, Abruf: 21.01.2015

Amazon Web Services (2015b): 1000 Genomes-Projekt und AWS, http://aws.amazon.com/de/1000genomes/, Abruf am 19.01.2015

Anderson, J. / Lehnardt, J. / Slater, N. (2015): Eventual Consistency, http://guide.couchdb.org/draft/consistency.html, Abruf am: 10.01.2015

Bertin-Mahieux, T. (2014): Million Song Dataset, http://labrosa.ee.columbia.edu/millionsong/, Abruf: 23.01.2015.

Bundesamt für Sicherheit in der Informationstechnik (o.J.): Cloud Computing Grundlagen, https://www.bsi.bund.de/DE/Themen/CloudComputing/Grundlagen/Grundlagen_node .html, Abruf am: 17.01.2015

Cloudant (2014): Pricing, https://cloudant.com/product/pricing/, Abruf am: 22.01.2015

Clustrix, Inc. (2014): Cloud Database and Database-as-a-Service (DBaaS) Market Projected to Grow, http://finance.yahoo.com/news/cloud-database-database-dbaas-market-130000399.html, Abruf am: 20.01.2015

Datenbanken Online Lexikon (2013a): Dokumentenorientierte Datenbank,http://wikis.gm.fh-koeln.de/wiki_db/Datenbanken/DokumentenorientierteDatenbank, Abruf am: 10.01.2015

Datenbanken Online Lexikon (2013b): Apache CouchDB,http://wikis.gm.fh-koeln.de/wiki_db/Datenbanken/CouchDB, Abruf am: 10.01.2015

Edlich, S. (2015):NOSQL Databases, http://nosql-database.org/, Abruf am: 10.01.2015

Gloster, F. (2014): Von Big Data reden aber Small Data meinen, http://www.computerwoche.de/a/von-big-data-reden-aber-small-data-meinen,3068465, Abruf am: 18.01.2015.

IBM (2007): The History of Notes and Domino, https://www.ibm.com/developerworks/lotus/library/ls-NDHistory/, Abruf am: 10.01.2015

IBM (2015): What is MapReduce?,http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/, Abruf am 20.01.2015

Izrailevsky, Y. (2011): The Netflix Tech Blog: NoSQL at Netflix, http://techblog.netflix.com/2011/01/nosql-at-netflix.html, Abruf am: 10.01.2015

Janssen, C. (o. J.): View,http://www.techopedia.com/definition/25126/view-databases, Abruf am: 10.01.2015

Jansen, R. (2010): CouchDB - angesagter Vertreter der "NoSQL"-Datenbanken, http://www.heise.de/developer/artikel/CouchDB-angesagter-Vertreter-der-NoSQL-Datenbanken-929070.html, Abruf am: 10.01.2015

Katz, D. (2005): What is Couch?,http://damienkatz.net/2005/12/what_is_couch.html, Abruf am: 10.01.2015

Klimt, W. (2013): NoSQL und Big Data: Was Sie über NoSQL wissen sollten, http://www.computerwoche.de/a/was-sie-ueber-nosql-wissen-sollten,2528753, Abruf am: 10.01.2015

Lennon, J. (2009): Exploring CouchDB, http://www.ibm.com/developerworks/opensource/library/os-couchdb/index.html, Abruf am: 10.01.2015

Salevan, S. (2011): MapReduce for the Masses: Zero to Hadoop in Five Minutes with Common Crawl, http://blog.commoncrawl.org/2011/12/mapreduce-for-the-masses/, Abruf am 20.12.2014.

ScaleDB (o.J.): Database-as-a-Service (DBaaS), http://www.scaledb.com/dbaas-database-as-a-service.php, Abruf am: 19.01.2015

The Apache Software Foundation (2014): 1.2. Why CouchDB? — Apache CouchDB 1.6 Documentation, http://docs.couchdb.org/en/1.6.1/intro/why.html, Abruf am: 10.01.2015

Walker-Morgan, D. (2010): NoSQL im Überblick, http://www.heise.de/open/artikel/NoSQL-im-Ueberblick-1012483.html, Abruf am: 10.01.2015

Wenk, A. (2014): CouchDB Introduction, https://cwiki.apache.org/confluence/display/COUCHDB/Introduction, Abruf am: 10.01.2015

Wikimedia Foundation (o.J.): Wikimedia Downloads, http://dumps.wikimedia.org, Abruf: 19.01.2015.



Konzepte und Einsatzszenarien von Wide Column Datenbanken

Concepts and Operational Scenarios of Wide Column Databases

Schriftliche Ausarbeitung
im Rahmen der Lehrveranstaltung „Integrationsseminar“
für das Kompetenzzentrum Open Source (KOS)

Vorgelegt von

Tim Hopp, Laura Maria Hoess
Nico Mueller, Linda Meier
Steffi Chan

am 26.01.2015

Fakultät Wirtschaft
Studiengang Wirtschaftsinformatik
WWI2012I

Table of Contents

List of figures

List of tables

1 Introduction

1.1 Problem statement

With the increasing amount of affordable technologies and ongoing digitalization of most areas of modern life, it became effortless to collect data on these areas to gain a deeper insight into the greater clockwork of our lives. However issues and challenges go along with this "explosion of data". The amount of data available is growing so rapidly that traditional storage systems, in particular database systems, cannot keep up processing these volumes. NoSQL database systems, especially wide-column databases, implement new ideas to natively support "Big Data". There are several systems available, some open source, some are to be purchased. Modern Organizations implement NoSQL wide-column databases to perform high performance operations, for example Google uses the Google BigTable to store indexed webpages and reply to user requests as fast as possible, for the Google search engine. But it is left to be examined what NoSQL wide-column systems are available for an open-source implementation to enable organizations to run tests for themselves and switch to this new approach. Additionally if these NoSQL-systems are as innovative as they claim to be.

1.2 Objectives

One of the main objectives of this paper is to provide a deeper insight on available and accessible, therefore open-source, NoSQL wide-column database systems, because of the overall scarcity of information available. To understand the capabilities of the examined systems, a list of criteria will be elaborated to provide a quick and comparable view on features and technologies. On basis of this list of criteria a suitable system will be chosen to be implemented as a prototype for later testing. The testing will consist of a comparison of the chosen NoSQL database prototype and a traditional SQL database in terms of implementation and basic performance specifications on the background of a fictional business case of logging system events. Furthermore the results will be evaluated and discussed to outline any differences between the NoSQL and SQL approach. Keeping in mind that the resources available for the research of this paper are quite restricted, a comparison to high performance, great scale test of End Point will be provided to finally make better predictions on the meaningfulness of the data raised by the prototype.

1.3 Structure

The paper is divided into four main parts: the introduction, theory necessary to understand the topic of NoSQL database systems, a practical part to discuss the decision for the prototype system, implementation, testing and ultimately result presentation and the conclusion to summarize the discoveries made. The theory focuses on the following topics:

- Big Data, to understand the need and most likely operational szenario of NoSQL wide-column systems
- SQL, to provide a quick overview of traditional database systems
- No SQL / wide-column, to introduce the topic and discuss the technology behind it
- Comparison of NoSQL and SQL systems, to outline familiarities and differences in the technologies

The following practical part of the paper will consist of:

- Wide-column database systems, to introduce several accessible solutions available
- List of criteria, to outline the features of the introduced systems and support a decision for the prototype
- Implementation of the prototypes
- Testing of prototypes, with several custom designed tests
- Comparison to the big scale End Point test

Finally the conclusion will provide an overview of the discoveries of this paper and an outlook on future developments in this field.

2 Basic Concepts of Databases

2.1 SQL

The Structured Query Language (SQL) was developed by IBM in the late 1970s. Successive versions then were endorsed as standards by the American National Standard Institute (ANSI) in 1986, 1989, 1992, 1999, 2003, 2006 and 2008. SQL has also been endorsed as a standard by the International Organization for Standardization (ISO).

SQL is not a complete programming language, rather it is a data sublanguage for creating and processing databases. To deliver the functions of a complete programming language, SQL needs to be embedded in scripting languages such as PHP or Java.

The Structured Query Language consists of three parts:

- Data Definition Language (DDL) for defining the data structure of a database
- Data Manipulation Language (DML) for inserting, deleting, reading and changing data
- Data Control Language (DCL) for user account control etc.

DDL and DML will now be explained. DCL is only of minor importance to the project and will be left out.[1]

DDL

With DDL tables and their structures can be created, changed and deleted. The basic commands, which are only used in DDL, are “CREATE” and “DROP”. Typical DDL statements look similar following:

```
CREATE TABLE name (
        attribut1 INTEGER,
        attribut2 VARCHAR(100),
        attribut3 VARCHAR(100),
        PRIMARY KEY (attribut1)
)

ALTER TABLE name
        ADD attribut3 DECIMAL(10,2),
        MODIFY attribut2 VARCHAR(50),
        DROP attribut3

DROP TABLE name
```

Furthermore, referential integrity constraints can be attached to each attribute of a table. Examples for constraints are:

- NULL
- NOT NULL
- DEFAULT ‘foo’
- CHECK (foo > 0)
- UNIQUE
- PRIMARY KEY (foo)
- FOREIGN KEY (foo1) REFERENCES (foo2)

[1] Kroenke, D.M./Auer, D.J. (2013), pp. 109 ff.

DML

As mentioned above, DML is used to create, read, update or delete (CRUD) data. The most important commands are INSERT INTO, SELECT, UPDATE and DELETE FROM. Here are some sample statements:

```
INSERT INTO tabelle (attribut1, attribut2) VALUES (5, 'Hans');
SELECT artikelgruppe, COUNT(*) AS anzahl
FROM artikel
GROUP BY artikelgruppe
HAVING COUNT(*) >= 4;

UPDATE tabelle SET attribut2 = 'Peter' WHERE attribut1 = 5;

DELETE FROM tabelle WHERE attribut1 = 5;
```

When manipulating data it is essential to include the WHERE key word. If the key word is left out, all tuples in a relation are manipulated or deleted.

2.2 NoSQL

A NoSQL (Not Only SQL) database is designed to handle a large amount of data which relational databases are not applicable of. The name emphasizes that those stores might support SQL query languages. A long time there was no alternative to relational databases but over the years more and more NoSQL databases were developed to fulfill todays' needs. These databases have become more popular in recent years because of a significant high data growth.[2] The data structures used by NoSQL databases are different from relational databases to increase their performance. They do not have a static schema anymore but a flexible structure that can be changed easily. Many of them are open-source projects even though the term "is frequently applied to closed-source systems". They run very well on clusters to support large volumes of data in comparison to relational databases.[3] Usually NoSQL databases also support automatic sharding, which means that data is automatically spread across a number of servers. In this way data is balanced across servers. If there is a case when a server goes down it can be quickly replaced. Many stores support replication to ensure high availability and a quick recovery. [4] A disadvantage of some databases are that they do not support ACID transactions. Overall, there are four different types of NoSQL systems

[2] Planet Cassandra (2015a)

[3] Sadalage, P.J/Fowler, M. (2012), p. 24

[4] mongoDB (2015)

which are key-value, document, column-family and graph which are described in more detail in the following paragraphs.

- Key-Value store
  - The Key-Value store is the least complex of all NoSQL databases. Each column consists of a key-value pair therefore it provides a two-dimensional quality.The first key is a row identifier for aggregation for the following columns. Some of them allow typing.[5]
  - Examples: Riak, BerkeleyDB
- Document store
  - This database is an extension of the key-value store. Each key is paired to a document that contains a complex data structure. They key has to be used to retrieve the document.[6]
  - Examples: MongoDB, CouchDB
- Column-family store
  - They are used for a large set of data and map keys to values and those values are grouped into multiple column families. It stores its content by column rather than by row.[7] Wide column stores can be seen as aggregated-oriented databases because often it is a two level aggregate structure.[8] More detailed information about the wide-column database types can be found in Chapter 2.5.
  - Examples: Cassandra, Hypertable, HBase
- Graph store
  - Data is stored in a graph store that can be easily presented in graph like networks.[9] It uses graph structures with nodes, edges and characteristics to store information.[10]
  - Examples: HyperGraphDB, Neo4J
- Multi-model store
  - In NoSQL there are some databases that include different data models which are known as multi-model store. Sqrrl, for example, include key-value, wide column, document and graph but is commercial project.

[5] mongoDB (2015)
[6] Planet Cassandra (2015a)
[7] Raj, P. (2014), p. 225
[8] Sadalage, P.J/Fowler, M. (2012), p. 24
[9] Planet Cassandra (2015a)
[10] Raj, P. (2014), p. 225

Ben Scofield rated every NoSQL data model in performance, scalability, flexibility, complexity and functionality. He also includes relational database. They should be considered when choosing the right database.[11]

Data Model	Performance	Scalability	Flexibility	Complexity	Functionality
Key–Value Store	high	high	high	none	variable (none)
Column-Oriented Store	high	high	moderate	low	minimal
Document-Oriented Store	high	variable (high)	high	low	variable (low)
Graph Database	variable	variable	high	high	graph theory
Relational Database	variable	variable	low	moderate	relational algebra

Table 1: Ben Scofield Rating

The rating shows that the performance of the key-value and column-oriented stores are high. For graph databases and relational database the performance is variable. The scalability is high for the key-value and column-oriented databases while the others are variable. In flexibility there are more differences because it is low for relational databases due to the fixed data structures they have. It is moderate for the Column-Oriented store because the schema has to be predefined. The complexity is moderate for relational databases, high for graph databases, low for document-oriented stores and column-oriented stores and there is no complexity for key-value store. The functional used for relational stores is relational algebra, for graph databases the graph theory is used and it is variable for document-oriented and key-value stores. It is minimal for column-oriented databases.

2.3 Comparison of SQL and NoSQL

The following chapter will provide an overview of the features of relational and NoSQL databases and summarize the different aspects of the two databases in a table.

First of all, relational databases can be used for different database operations[12] for example to store, manage or retrieve data through applications and queries. In general, these databases consist of tables[13]. Data is stored in columns and rows, therefore the user needs to

[11] Scofield, B. (2010)
[12] Bhattacharjee (2014)
[13] Linwood, J./Minter, D. (2010), p. 183

know what to store in advance.[14] In addition to that, a relational database scales its data vertically and has fixed relationships.[15]The sizes of rows and columns are fixed.



[14] mongoDB (2015)
[15] Linwood, J./Minter, D. (2010), p. 183

In comparison NoSQL databases do not have a fixed structure of data, it is stored dynamically. Furthermore, relationships are not enforced and the Data types for a given attribute are not consistent. Typically, NoSQL databases do not fully support ACID transactions[16], but they support the needs of today`s modern business application. Massive scalability or flexibility[17] is one of the most important criteria of those databases which have become the first alternative to relational databases.[18] They also can handle large volumes of data[19], are very efficient and support agile sprints. Like other database, NoSQL may have some disadvantages in some points, like the fact that there is no mandated structure. Therefore, migrating an application to NoSQL may be a problem[20].

To get a better overview of the differences between a relational and NoSQL databases, the following table shell compare the most important facts.[21]

	Relational Databases	**NoSQL Databases**
Types	One type (SQL database)	Many different types including key-value stores, document databases, wide-column stores and grap databases
Development History	Developed in 1970s to deal with first wave of data storage applications	Developed in 2000s to deal with limitations of SQL databases, particularly concerning scale, replication and unstructured data storage
Examples	MySQL, Oracle Database	Cassandra, HBase, MongoDB

[16] Kuznetsov, S.D./Poskonin A.V. (2014), p. 323

[17] Linwood, J./Minter, D. (2010), p. 183

[18] Planet Cassandra (2015b)

[19] mongoDB (2015)

[20] Linwood, J./Minter, D. (2010), p. 183

[21] With modifications taken from: mongoDB (2015)

Data Storage Model	Each column has a specific information about what needs to be stored (e.g. “car”, “owner”, “brand”, etc.) and individual records are stored as rows. A column predefines the data type and the data have to fulfill it. More than one table can be joined together to select the needed data. For example, “owner” might be stored in one table, and “car” in another. When a user wants to find more information about the owner, the database engine joins both tables together to give the information necessary.	It depends on the database type. Cassandra for example has two columns which are “key” and “value”. The first key is often described as a row identifier for aggregation of the following columns. Document databases store all relevant data together in a single “document” in JSON, XML, or another format, which can next values hierarchically.
Schemas	Structure and data types are fixed in advance. To store information about a new data item, the entire database must be altered, during which time the database must be taken offline.	It usually is dynamic. Records can add new information on the fly, and unlike SQL table rows, dissimilar data can be stored together as necessary. For some databases (e.g. wide-column stores), it is somewhat more challenging to add new fields dynamically.
Scaling	Vertically, meaning a single server must be increasingly powerful in order to deal with increased demand. It is possible to spread SQL databases over many servers, but significant additional engineering is generally required.	Horizontally, meaning that to add capacity, a database administrator can simply add more commodity, servers or cloud instances. The database automatically spreads data across servers as necessary.
Development Model	Mix of open-source (MySQL) and closed source (Oracle Database)	Open-Source
Supports Transactions	Yes, updates can be configured to complete entirely or not at all	In certain circumstances and at certain levels (e.g. documents level vs. database level)
Data Manipulation	Specific language using Select, Insert and Update statements	Through object-oriented APIs
Consistency	Can be configured for strong consistency	Depends on the product. Some provide strong consistency (e.g. MongoDB) whereas others offer eventual consistency (e.g. Cassandra)

Table 2: Comparison of Relational Databases and NoSQL Databases

In conclusion one can say that it is hard to decide which database is superior. NoSQL provides lots of advantages, these kind of databases also have got their own challenges and strengths. It depends on the specific use case[22] which database to adopt.

2.4 Relational Database Solutions

The relational model, developed and published in 1970 by Edgar Frank Codd, is a database model based on relational algebra for tracking information about things, formally called entities. Key terms of the relational model are relations, tuples and attributes. This set of terms can be compared to traditional data processing, which uses files, records and fields, or mathematics, which uses tables, rows and columns. A database compliant with the relational model consists of a set of separate relations, in which each relation contains information about one and only one kind of entities ("theme").

Every relation is a table, but not every table is a relation. Table 1 shows the criteria for a relation:

Criteria for a relation
Rows contain data about an entity
Columns contain data about attributes of an entity
Cells of the table hold a single value
All entries in a column are of the same kind
Each column has a unique name
The order of the columns is unimportant
The order of the rows is unimportant
No two rows may hold identical sets of data values

Table 3: Criteria for a relation

2.4.1 Normalization

Big tables usually contain more than one theme and thus they hold information about multiple entities per row. This is why the core principle of relational data modeling is normalization, which means breaking down big tables into smaller ones to avoid redundancy and to turn it into relations. Therefore, each tuple is assigned a primary key, which can be stored in tuples of other relations as a foreign key.

[22] Linwood, J./Minter, D. (2010), p. 183

Information about a particular entity then is distributed over several relations, but the parts are unambiguously connected via foreign keys. When requesting information about an entity, SQL can reassemble the parts by joining the relations. The steps of normalization are the following:

Normalization
1. Identify all the candidate keys of the relation (determine all the other attributes!) 2. Identify all the functional dependencies in the relation. There may be determinants which do not identify all attributes of the relation but only a part. 3. If a determinant is not a candidate key, the relation is not well-formed. In this case: a. Place the columns of the functional dependency in a new relation of their own b. Make the determinant of the functional dependency the primary key of the new relation c. Leave a copy of the determinant as a foreign key in the original relation d. Create a **referential integrity** constraint between the original relation and the new relation 4. Repeat until every relation is well-formed

Table 4: Normalization

In order to understand the normalization procedure, the terms 'functional dependencies' and 'keys' need to be explained.

Functional dependencies

Functional dependencies tell which attribute is dependent on another. Here is a simple example:

- Total = Quantity * $5
- Quantity → Total

In this example, we say that "Total" is functional dependent on "Quantity" or "Quantity" is determined by "Total". Therefore, "Total" is a determinant. If you change "Quantity" you will get another "Total". This also applies to non-mathematical correlations. For example "ID" usually determines "username". For each ID there is a specific username.

Keys

A key is one or more columns of a relation that is used to identify a row. There are different kinds of keys:

- A candidate key is a key of a relation that functionally determines all the other attributes of the relation.
- A primary key is the candidate key chosen by the developer to uniquely represent each tuple of the relation.
- A foreign key is a primary key of one relation that is placed in another relation.[23]

2.4.2 Concurrency Control

Concurrency Control ensures that one user's work does not influence another user's work. In relational databases this is managed by atomic transactions. An atomic transaction is a series of actions to be taken on a database such that all of them are performed successfully or none of them are performed at all.

Two major consistency levels are called 'statement' and 'transaction'. A statement is a single command (read, write or change data), whereas a transaction is a bundle of statements. Statement-level consistency means that during a statement required relations are locked so that only a single user can access the data. However, if a user executes a bundle of statements, the required relations will be unlocked between the statements and other users may access and change data. With transaction-level consistency required relations are locked for the entire transaction. RDBMS usually support transaction-level consistency.[24]

2.4.3 Final Words on RDBMS and SQL

SQL and RDBMS are strongly intertwined and can hardly be discussed independently. SQL is a rich language with complicated subqueries, but the data model can only handle simple features. That means, in order to develop an SQL database design, people usually develop an entity-relationship model first. But then, several steps will transform the ER model to a relation-model. After the transformation, many features of the ER model like aggregates and subtypes have been replaced by simple foreign key relationships. Although it is possible to reconstruct the features, the RDBMS is not aware of them and cannot take advantage from that knowledge.

The core procedure in building a relational database is normalization, in which all redundancies are removed from an SQL database. This causes that information about a single entity may be distributed over several tables. With SQL the parts can then be put together again.

[23] Kroenke, D.M./Auer, D.J. (2013), pp. 59 ff.
[24] Kroenke, D.M./Auer, D.J. (2013, pp. 306 ff.

2.5 Big Data

Being certainly one of the most discussed and interesting topics in the IT business, it is clear that many people do not even know what exactly the term means. The Google Trend charts show clearly that the interest in Big Data has risen, especially over the last two years. So everyone talks about it, but no one knows what it is. The following chapter will discuss the term big data and related topics to understand the idea and need behind wide column stores, as discussed in this paper.

Definition Big Data

When confronted with the term of Big Data, one simple explanation should come to mind: It represents a set of an immense volume of data. However this short definition does not quite capture the whole meaning of the term and the connected field of business and research. To define Big Data more precise, literature offers different approaches to the field as follows:

Edd Dumbill defines Big Data as data that exceeds the processing capacity because of volume, its fast-moving nature and mismatches the structure of your database.[25] The focus is once again set on the sole data aspect once again. A more differentiated position is the following:

“Extracting insight from an immense volume, variety and velocity of data, in context, beyond what was previously possible.”[26]

This definition by Dirk DeRoos introduces a super-level of Big Data, the gain of information extracted from high volume data sets. This is an important aspect of the field, which will be explained more detailed later on. However the definition lacks the connection of the pure data/physical aspect of Big Data and the more purpose oriented one mentioned here. Another possible approach is:

[25] DeRoos, D. (2012)
[26] Dumbill, E. (2012)

“Big Data is a phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex, and diverse types of data. Big Data is often defined along three dimensions -- volume, velocity, and variety. […] Addressing the challenge and capturing the opportunity requires advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.”[27]

By the standards of the arguments made above this definition offers a well-rounded explanation, considering both the physical and meta-physical approach to Big Data.

Origins of Big Data

Along with Big Data comes the term of the information explosion, in other words the vast amount data which became available over the past few years. The following factors led to the emergence of Big Data:

- Availabilty of new technologies:
    - New technologies are able to capture more data. Additionally the steady decline in prices for technologies ensure that a many people are able to afford gadgets, which enables a wide base of data sources. For example there are 4.6 billion cell phone subscribers worldwide.
- Ease of share
    - With the upgrade of technology the power of the internet increases significantly. It offers users the opportunity to easily share data. According to Cisco Systems the annual flow of traffic over the internet reached 667 exabytes and it continues to grow.[28]

Examples of Big Data

Popular examples of Big Data can be found in several large companies, as they tend to handle large amounts of data in their day-to-day business. Two examples of well-known organizations follow.

Walmart, one of the biggest retail cooperation of the world, processes more than 1 million customer transactions every hour alone. The estimated size of the database absorbing the data is more than 2.5 petabytes, which would be the equivalent of 167 times the amount of books in the American Library of Congress.

Another example the operator of the biggest social-networking website, Facebook. The estimated amount of photos on Facebook resolves around 40 billion. Assuming that Facebook

[27] TechAmerica Foundation (w. y.a)
[28] The Economist (2010)

compresses the file size of the photos in its database to the minimum amount of 50 kilobytes per picture, the estimated overall size of said database would be 2000 petabytes of data for photos alone.[29]

Opportunities

Considering the examples given above, it becomes apparent that the volume of data is vast in current businesses and grows rapidly. Due to the increasing degree of digitalization of many areas of everyday life, it is simply easier to obtain these amounts of data. This enables possibilities that could not be ceased in earlier times, for example detect business trends and prevent diseases or even fight crime and many more. Additionally it can discover and unlock new sources of economic value, for example by monitoring customer purchases and analyzing them, or promote new scientific discoveries by considering a larger data set, which enables deeper insight.[30]

Problems

The high availability of new data creates new challenges as well. New technology enables the capturing, processing and sharing of large amounts of data, but it gets increasingly difficult to simply store it. The chart below plots the availability of data against the available storage space to visualize the problem.

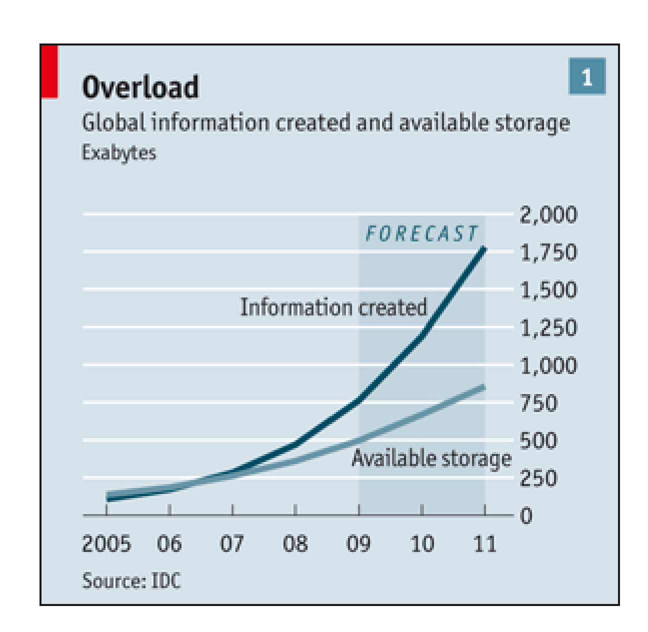


Figure 1: Overload[31]

[29] The Economist (2010)
[30] The Economist (2010)
[31] Contained in: The Economist (w. y.a)

Moreover, Big Data demands for new solutions in areas like data security, privacy protection and the detection and handling of redundant data.[32]

The relation of Big Data and Wide-Column Stores

One of the main reasons of for growing popularity of NoSQL-databses in modern businesses is to tackle the issue of big data. Big Data, like mentioned above, is marked by data velocity, variety, volume and complexity. NoSQL systems have several design features that specifically serve these characteristics.

- Continuous data availability
  - NoSQL databases are designed to operate on distributed nodes, which ensures minimal data redundancy and high availability. If a single node crashes, another node can simply continue operations. Another advantage of the distributed nature of NoSQL systems is the high degree of scalability. It is quite simple to exchange or set up additional nodes if the necessity arises.
- Real location independence
  - NoSQL databases are able to write or read data independently of the node that responds to an I/O request. That reduces the workload on each individual node, which mainly servers the purpose of handling high volumes of data.
- Flexible data models
  - NoSQL database do not rely on traditional predetermined models of data. This high degree of flexibility enables the system to accept all kinds and forms of data, regardless if it is structured, semi-structured or even unstructured.
- Simple analytics and business intelligence mechanisms
  - High volumes of data can provide a more differentiated insight on businesses. Modern NoSQL solutions provide not only the architecture to handle high amounts of data, but deliver already integrated data analytics, which offers a fast insight on large data and accelerates the decision-making process. This feature also obsoletes the need for additional business analytics software, because of its ability to serve most of the basic analytics requirements.[33]

[32] The Economist (2010)

[33] Cf. Planet Cassandra (2015a)

2.6 Wide Column Database Solutions

A wide column store is one type of NoSQL database.[34] It can be viewed as a two-dimensional Key-Value-Store. Of cause it is also possible to implement column-stores with SQL. This chapter will provide a brief overview on their advantages and the reason why there is a permanently increasing interest in them in the field of analytics.

2.6.1 Row-oriented vs. Column-oriented

It is easiest to explain these concepts on the basis of a comparison. Figure 2 shows a very simple to understand example - A sales database.

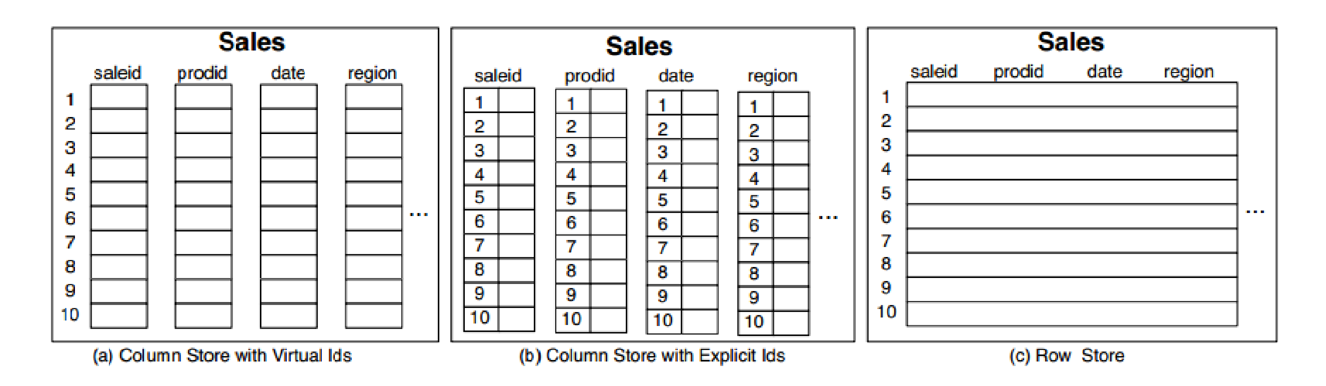


Figure 2: Physical layout of column-oriented vs. row-oriented databases[35]

Assume a query checks how many sales were made in each region displayed. A row oriented database (c) always has to retrieve the whole row – saleid, prodid, data and region. A column oriented database has the ability to only read the corresponding columns – saleid and region. Under these circumstances it does not make a big difference in the example table, but in databases with countless columns, a row-store would certainly perform awful. A column-oriented database on the other hand would work with fractions of each row and would therefore not outperform the row-orientated model greatly. It could read large amounts of data much quicker. These databases can also be referred to as column-stores, wide column stores or wide column databases. Column-stores (a) (b) have a complete vertical partition and write on disk in data blocks. Each column is stored together. [36] Basically a column-store is a collection of separate columns in which each and every one of the columns can be accessed individually. This approach helps to use CPU potential, optimize in- and output and use bandwidth much more efficiently than they are used in row-stores. Since databases today become more and more complex and hold more and more information, row-oriented databases struggle to meet good performance ratios. Database performance and velocity

[34] Cf. MongoDB (w.y.)
[35] Abadi (2012), p. 199
[36] Cf. Jhilam (2011)

may be seen as the most significant bottleneck today. Therefore column-stores perform due to their architecture much better in these ratios.[37] The focus in this paper will be set on wide-column databases, which means having huge amounts of attributes in a database. Literature also discusses the use of column-stores for sparse data. But due to the scope of this thesis these discussions are not relevant.

2.6.2 Data model

The data model was shown briefly in Figure 2 (a) and (b). Column-stores may be distinguished by either having virtual IDs or explicit IDs. Basically explicit IDs, the data model of key-value-stores is the background set up of wide column databases. Virtual IDs like in Figure 2 (a) are visible to the user. The user is not required to know where each column is stored, but needs to know the keys of each rows. Figure 3, which shows the Cassandra data model, demonstrates this approach in detail.

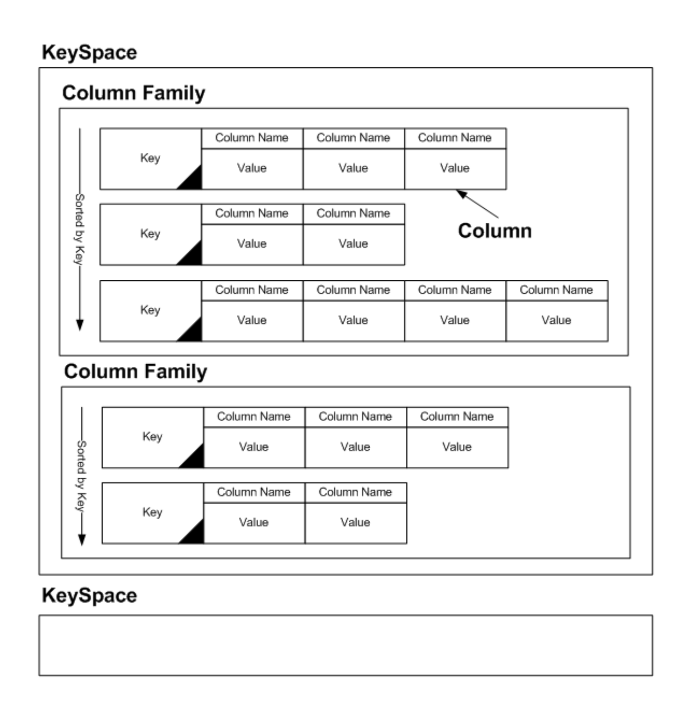


Figure 3: Cassandra Data Model[38]

The model shows KeySpaces, which represent individual, logical databases. Each KeySpace consist of Column Families which again consist of rows of columns. Each column, like in Figure 2 (a) has a corresponding Key to clarify correspondence of columns. Figure 3 shows that this approach enables certain columns not to exist in every row, because wide-column stores allow theoretically for an infinite number of columns. Additionally each value is often stored with a corresponding timestamp.

[37] Cf. Abadi (2007), p. 9
[38] Cho (2010)

2.6.3 Functionalities

The features of wide-column-stores have in many cases been inspired by research and features of row-store systems.

In databases where every entry is stored in an array with fixed length, the position of data can easily be calculated by an algorithm like the following – start of column X + xth value*width of column.[39] But this would enlarge the database greatly the more data is handled. Many modern databases therefore implement compression algorithms. The compression feature can achieve a considerable reduction of the overall data size and the amount consumed on the disk. Because of compression, for example by implementing arrays with non-fixed-length, databases need then more sophisticated algorithms to locate data entries.

Since NoSQL databases lack functionalities like GroupBy, Join or any other arithmetic operations, in many cases they attempt to overcome this problem by implementing the MapReduce concept. Fairly simplified it consist of the three stages: map, shuffle and reduce. This concept is often implemented by the Hadoop file system, which simplifies the complexity of output shown to the user. A user can write map functions, which transform data, and reduce functions, which aggregate data, to work with the data inside a wide column database. To understand MapReduce better the following example will explain the process.

MapReduce could be compared to sorting cards. The intended result may be to sort all numeric cards by suit and then calculate the total value of each suit. In the beginning one stack of cards is available, just like a data input would. In the map phase then this data will be taken and divided it into individual records. This is comparable to the outlay of the card deck. The following phases are the shuffle and sort phase, which means sorting the card deck by each suit and getting rid of kings, queens, jacks and jokers. Once the sort is done, the reduce phase starts. In this example it would be the summation of the numeric cards. The four calculated sums would be the output. [40] Figure 4 shows the process.

[39] Cf. Abadi (2007), p. 39 seqq.
[40] Cf. Anderson (2013)

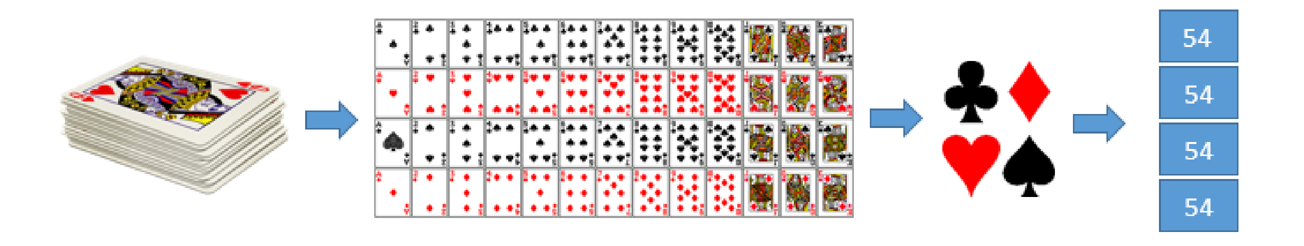


Figure 4: MapReduce Example: Input, Map, Shuffle & Sort, Reduce and Output

Obviously the MapReduce process more suitable for huge amounts of data when executed on a multi-node system, because a single job can be broken up into smaller parts, where each node only completes a fraction of the overall task. Once each node completed its task, the output of each individual node is consolidated again. Figure 5 shows a more complex view of the process.

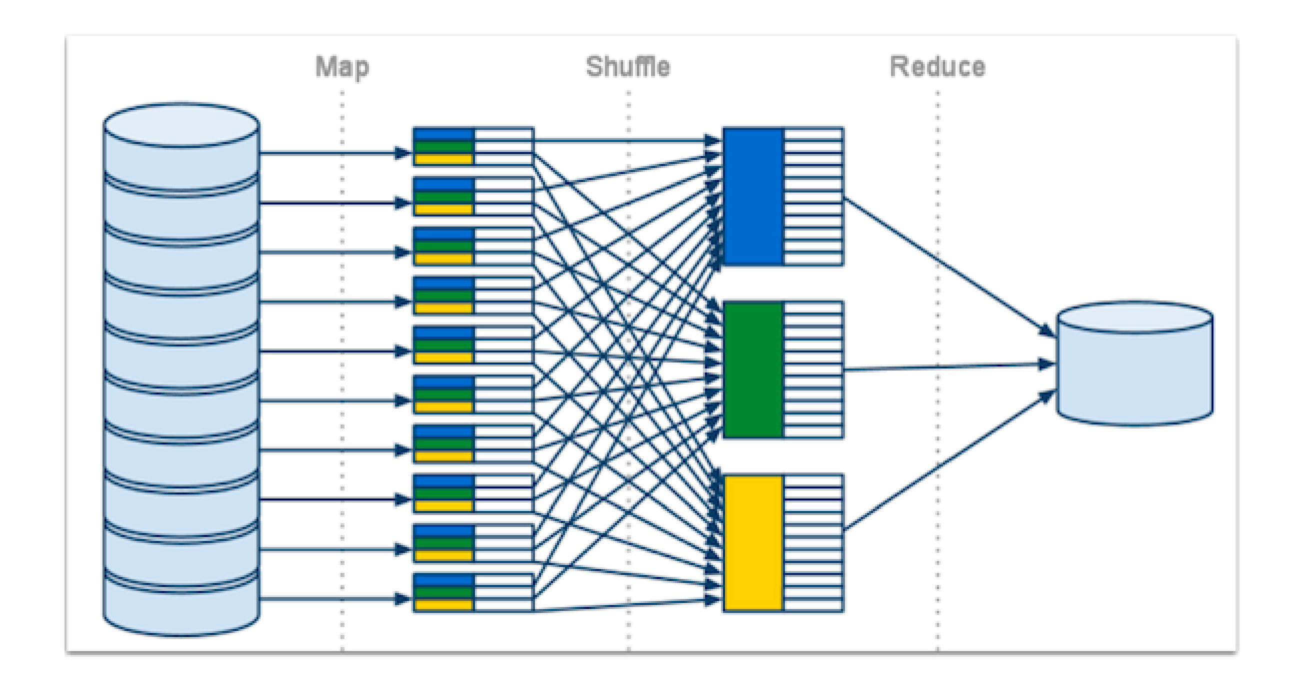


Figure 5: Theoretical view of MapReduce[41]

[41] Google Cloud Platform (w.y.)

3 Practical part

3.1 Description of Database Products

3.1.1 Cassandra

Cassandra is a top level Apache project. Initially, Cassandra was developed at Facebook and based on Amazon’s Dynamo and Google’s BigTable. The database product is designed to run as clusters of hundreds of nodes. As the clusters are decentralized, there is no single point of failure. There are no master/slave structures, especially there is no “original dataset” and its “backup” – all datasets are equally important, even if they are replicas. Clients can also contact any node in a cluster which then will coordinate the request. The nodes of a cluster can be placed in multiple geographic areas or even partly or entirely in the cloud. Cassandra is also aware of the locations of the several nodes and uses this knowledge to enable native multi data center backup and recovery.

Cassandra has been designed for high write throughput while not sacrificing read efficiency. This fits the big data requirements where huge amounts of data are constantly fed into the database. Some of the most impressive performance metrics are Cassandra’s sub-second respond times and linear scalability.[42]

The underlying data model is simple and dynamically accommodates changes to data structures. Cassandra uses an SQL-like query language called Cassandra Query Language (CQL) offering almost all SQL-features with the exception of features like stored procedures and foreign keys.[43]

This chapter will first introduce the database internals of typical write and read requests from a client view. Then a write request on a single node will be discussed with respect to its file system and durability. Next, Cassandras features will be compared to the ACID-model, which is a standard for SQL-databases. Finally, a short use case will demonstrate Cassandra’s simplicity and its enormous performance advantages over traditional databases.

[42] Planet Cassandra (2015b), p. 1
[43] Datastax (2015a), pp. 6 ff.

Writing to the Cluster

A write request (also known as “write mutation” or simply “write”) can be an INSERT, UPDATE or DELETE command. Clients can send writes to any node of a cluster, which then becomes coordinator of the request. Depending on partitioning and replication settings the coordinator will forward the request to specific other nodes, where it is processed and safely stored (for more information see Durability). Each node then sends a positive answer back to the coordinator, which then sends a positive answer to the client. If one or more of the nodes send a negative answer, the coordinator will check consistency before giving a positive answer to the client.[44] Two major settings influence the decision:

- Replication Factor
- Consistency-Level

The Replication factor simply sets how often one row needs to be represented in the cluster. A replication factor of 3 means that each row needs to be stored on three different nodes.[45]

The Consistency-Level determines how many of the replication nodes have to send a positive answer before the entire write process can be considered successful and the coordinator can acknowledge the write to the client. Each request can be given another consistency-level.[46] Some of the available options are:

Level	Description	Required number of positive answers with a replication factor of 3
ALL	All replica nodes must send a positive answer	3
LOCAL_QUORUM	A quorum of the replica nodes in the local data center must send a positive answer	2
EACH_QUORUM	A quorum of all replication nodes in all data centers must send a positive answer	2
ANY	One replication node must send a positive answer	1

Table 5: Write-Consistency Levels[47]

[44] Planet Cassandra (2015c), pp. 2 ff.
[45] Datastax (2015b), p. 13
[46] Datastax (2015b), p. 65
[47] Datastax (2015b), pp. 65, 66

Assume that a cluster's replication factor is 3 and there is a write request using a consistency-level of QUORUM. The coordinator (A) saves the write in its own system and send a request to two other nodes (B and C) for replication. Node B sends a positive answer but node C seems to be down. As the quorum of 3 is 2, consistency-requirements are met and the coordinator acknowledges the write request to the client. As consistency can be met, the coordinator leaves a hint on any node (e.g. "D") containing the location of the downed node C and the original request. Once C comes up again, D remembers that it is still holding a hint and sends the original write request to C. C then processes the request and full replication is restored. For the case that consistency cannot be met, the coordinator will not send a hint.

Be aware that B does not roll-back the successful write request just because full replication cannot be achieved. Each cell gets a timestamp of the latest update and so it does not pose a problem that A and B hold other/newer information of the same row as C.[48]

Reading from the Cluster

Also for read requests ("reads") a client can contact any node of a cluster, which becomes the coordinator of the read request. Based on partitioning and replication settings the coordinator will know on which nodes the requested rows can be found and send read requests to these nodes[49]. Each row will be treated atomically, but not the entire read request. Each requested row will undergo the same process as writes. Using replication factor and consistency-level the coordinator collects positive answers from the replica nodes. If enough positive answers have been received, the coordinator sends a positive answer to the client. If the replica nodes send different data about the same row to the coordinator, the coordinator picks the most recent timestamp.

Writing to a Node

When a write request is sent to a node, the node stores the request in two locations: a MemTable in memory and a CommitLog on disk. Both MemTables and CommitLogs are maintained separately for each table on the node. At given time a MemTable will be flushed to an SSTables on disk. SSTables (Sorted Strings Table) are append-only file structures. Instead of overwriting existing rows in an SSTable Cassandra appends a new version of the row with a newer timestamp. Periodically multiple SSTables of one CQL-table are consolidated using only the newest version of each row. As soon as a request is flushed from a MemTable to an SSTable, the corresponding entry in the CommitLog will be erased. The CommitLog is not more than a safety-copy of the MemTable on disk.

[48] Datastax (2015b), p. 63
[49] Datastax (2015b), p. 63

It is not used for writing into SSTables as disk is slower than memory. But if the node unexpectedly powers off and memory clears, the CommitLog still holds all requests that are pending for execution. Once the node comes up again, the MemTables are repopulated using the CommitLog's information.[50]

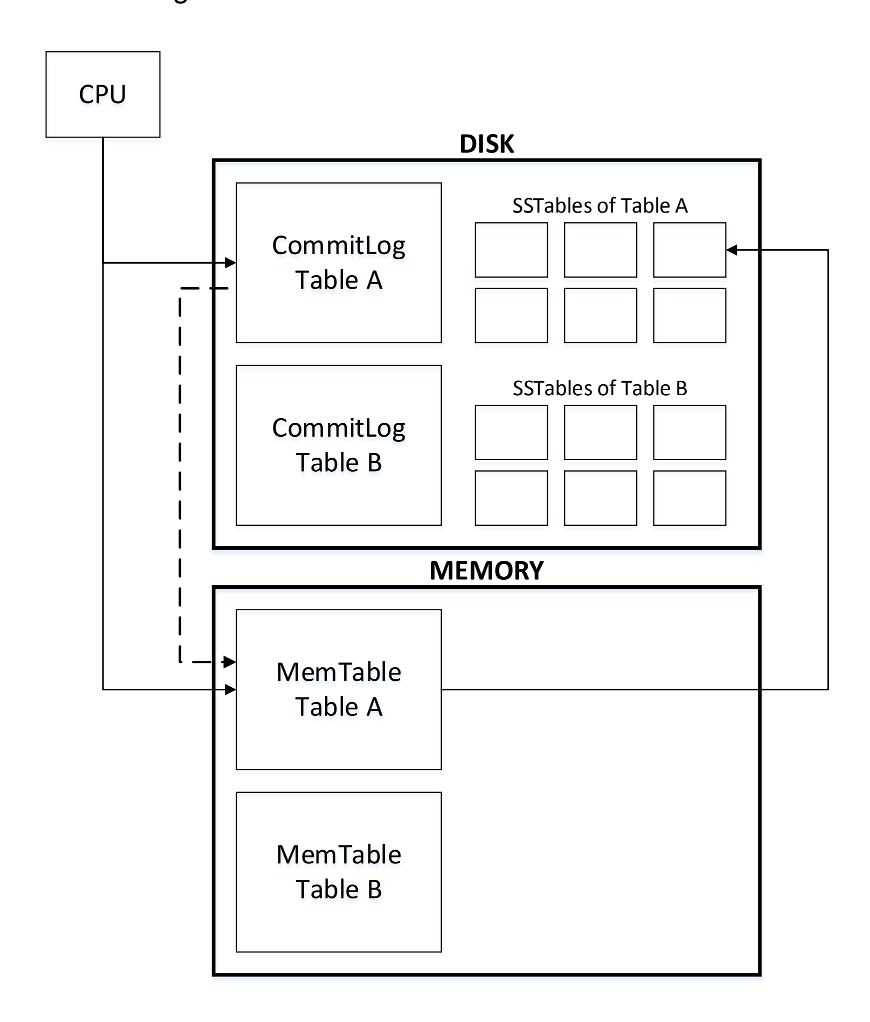


Figure 6: Writing to a single node[51]

Atomicity

Cassandra provides partition-level atomicity, which means that inserting or updating one row on one node is treated as one operation. As there are no foreign keys in Cassandra, multiple requests do not need to be bundled in one transaction and SQL-transactions containing multiple requests are not supported.

[50] Planet Cassandra (2015c), pp. 2 f.
[51] Planet Cassandra (2015c), p. 2

Consistency

As indicated above Cassandra uses tunable consistency. For each request the user or client application can decide which consistency-level is used. If a consistency-level smaller than ALL is selected, performance can be increased. Each row of each table is replicated on a different set of nodes and only row-level isolation is in place, which means that there are no locks when concurrently updating multiple rows and tables. That means that when trying to replicate a write request some of the replica nodes may be too busy to accept the request. Assume we have a replication factor of 5 and a consistency-level of QUORUM. A request is replicated to 5 nodes and 2 of them are busy. Anyways, the coordinator receives positive answers from the other 3 nodes and acknowledges the request. The client receives an answer very fast because he does not have to wait for all 5 nodes. As soon as the nodes become available again, replication can still be restored. This, of course, is not consistency in the classical ACID-manner. But at the end of the day Cassandra's numerous mechanisms will ensure that consistency is met anyways with the plus of an increased performance.

Isolation

Row-level isolation is in place. That means that a row which is being written to by one client cannot be seen by any other client until the request is completed.

Durability

To ensure durability, each node will acknowledge a write request if, and only if, it has been written to MemTable and CommitLog. Changing the settings the node will even wait with a positive answer until the request has been flushed from the MemTable to an SSTable.[52]

Use Case: Workload Separation

A company has three data centers: main, analytics, search. The configured replication strategy states that each row needs to be saved on 3 nodes in the main data center, 2 in analytics and 1 in search. Load-balancing and proxy server settings ensure that client requests will always be sent to the main data center. If the consistency-level is LOCAL_QUORUM or less, the coordinator node in main data center will acknowledge a write request as soon as all 3 nodes in main data center hold replicas of the request. In parallel, the coordinator sends more replicas to analytics and search. While the replicas are traveling through the slower WAN-connections to the other data centers, the client has already got his sub-second response. Shortly after that the replicas will also be written to the nodes in analytics and search.

[52] Datastax (2015b), p. 5

Analytics gets all information instantly without a complicated ETL process and Search also gets updated immediately without waiting for data center synchronization. As a plus, the company gets a geo-distributed live-backup using three different locations.[53]

3.1.2 HBase

Apache HBase is an open-source and distributed database, modeled with Google's Big Table in mind and written in Java.[54] It is designed "to manage extremely large data sets".[55] HBase has been developed since 2007 and was part of Apache Hadoop.

Basic information

Today HBase is available under the Apache License 2.0. Shortly after the release of Google's Big Table in November 2006 an initial HBase prototype was created and only a few months later in October 2007 the first "usable" HBase was developed. NoSQL databases got very popular, so in May 2010 HBase became an Apache top-level project. Before that it was only an Apache subproject.[56] HBase is more a data store than a database because it lacks many features that can be found in a relational database management system such as typed columns, secondary indexes, transactions and advanced query languages. This database is suitable to store hundreds of millions or billions of rows that is why enough hardware is needed so HBase can run quite well.[57]

Features

A of lot features come with HBase which are listed below:

- HBase is a consistent database with strong reads and writes means that HBase is not "eventually consistent" like other NoSQL databases.
- A lot of supported programming languages: C, C#, C++, Groovy, Java, PHP, Python and Scala
- The maximum value size is 2TB
- HBase supports automatic sharding
- Automatic RegionServer failover
- HBase supports MapReduce and stored procedures
- Easy access methods: HBase supports Java API
- Thrift/REST API: HBase supports Thrift and REST for non-Java Front-ends

[53] Planet Cassandra (2015b), pp. 63 f.
[54] Apache HBase (2015a)
[55] Grehan, R. (2015)
[56] George, L. (2011)
[57] Apache HBase (2015b)

- HBase is built on top of Hadoop Distributed File System so it provides fast record lookups and updates for large tables[58]

The official website has a lot of useful documentation which includes guides, videos and presentations about HBase. It is a popular product and very similar to Google's Big Table. The supported server operating systems are Linux, UNIX and Windows.[59]

The integrity model is the log replication and atomicity, consistency, durability are supported. There is a role-based access control to restrict certain functions. In addition to that, the replication can be set to master-slave, master-master and cyclic. For backup and recovery the user has different options to choose from. A few of them are export, CopyTable and Cluster Replication. It is possible to export the table with Export MapReduce, to copy the table data into a Sequence file on HDFS. If another HBase cluster is available that the user can treat as backup cluster it is recommended to use the tool CopyTable to copy a table at a time. A third options is the Cluster Replication. The backup cluster does not have be identical with the master cluster so the backup cluster can be less powerful therefore cheaper while still having enough storage for a backup.[60]

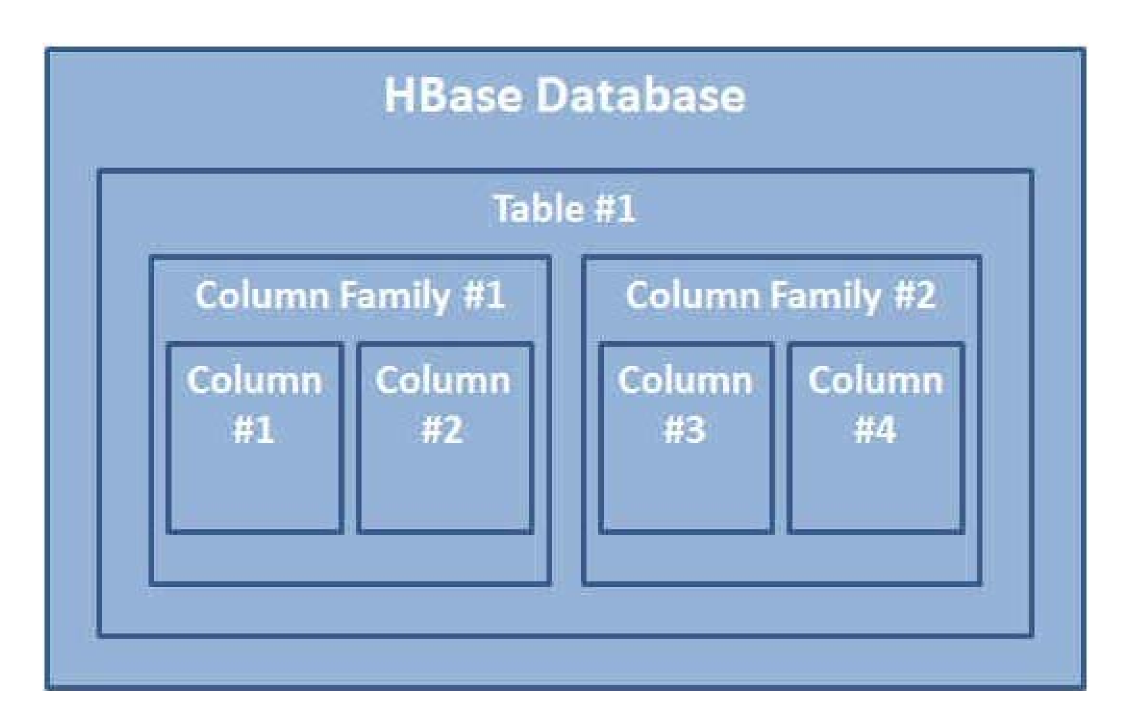


Figure 7: HBase Database Structure[61]

An HBase system has a set of tables and each table contains rows and columns like a relational database. Each table has an element which is defined as a Primary Key and all access

[58] Apache HBase (2015a)
[59] Apache HBase (2015b)
[60] Sematext (2011)
[61] With modifications from: Toad World (w. y.a)

attempts to HBase tables have to use this Primary Key. An HBase column has an attribute of an object that means the column is always related to the table. For example, the table is cars and each row stores information about that car, a typical column might be the car brand to specify it. Many attributes can be group together into column families so “that elements of a column family are all stored together”. The user needs to predefine the table schema and the column families. It is still flexible because new columns can be added to column families at any time.[62]

3.1.3 Hypertable

Hypertable is an open source database system based on Hadoop, which was released in 2009 by Hypertable Inc.and relies on NoSQL. It was strongly influenced by the design and properties of Google`s Big Table.[63] According to IBM it therefore has a “proven scalable design that powers hundreds of Google`s services”. [64] The main focus of Hypertable is to work with high volumes of data, just as Google BigTable does.[65] A documentation is located on the official Hypertable website.



Figure 8: Hypertable logo

To achieve higher performance Hypertable is almost entirely written in C++. Although it still supports various applications in other programming languages like Java, PHP, Phyton, Perl and Ruby. Such applications can access the database via the C++ API Apache Thrift, in the terminology of Hypertable called ThriftBrokers. Supported server operating systems incorporate Linux, Mac OS X and Windows. Mentionable is that Hypertable provides all ACID components. This includes atomicity, consistency, isolation, durability and concurrency levels. Multiversion Concurrency Control (MVCC) is provided as the Integrity Model, it ensures that user access is at any time possible. Notable is that Hypertable can not only run on top of the Apache Hadoop DFS, but also on GlusterFS or Kosmos File System (KFS).

Designed to overcome scalability problems, Hypertable is one of the five wide column databases listed on the db-engines website. On the site it scores very low and is far behind the

[62] IBM Corporation (2014)
[63] Cf. Hypertable Inc. (w.y.)
[64] Cf. Hypertable Inc. (w.y.)
[65] Cf. Scale Out with Hypertable (w.y.)

most popular wide column databases – Cassandra and HBase. [66] In contrast to other databases which often use hash tables, Hypertable identifies cells by a four-parted key. It stores the row as string, column family as byte, the column qualifier as String and the timestamp as long integer. This additionally enables the database to store different cell versions.

As a wide column database, Hypertable does not support typical data types, but uses opaque byte sequences. Also it neither supports joins, nor transactions, because both would lead to tremendous loss of performance when dealing with e.g. Petabyte-sized tables. In addition the high performance and reduced request latency leads to higher application throughput and quicker operations. [67] Figure 9 provides a brief overview of the database.

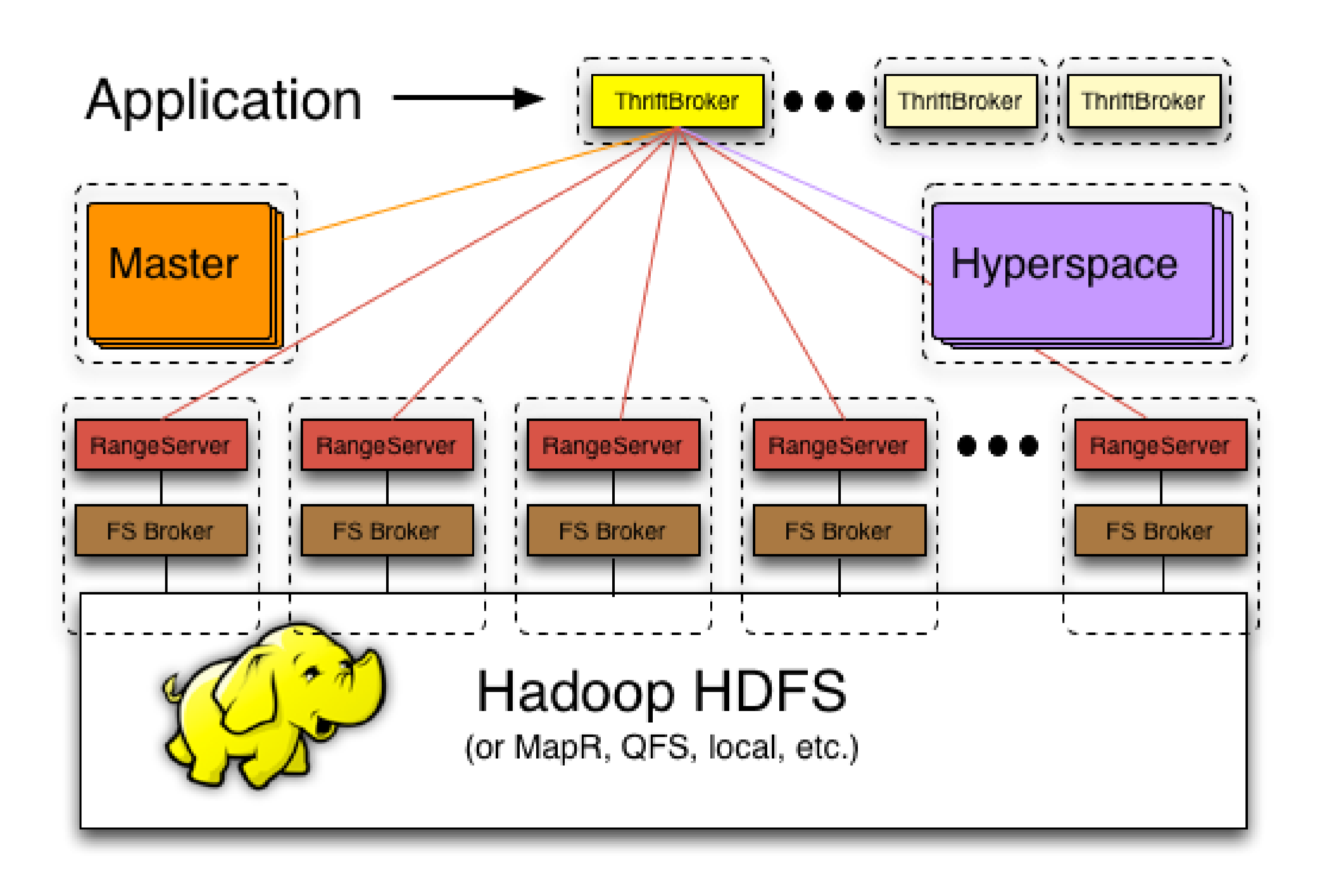


Figure 9: Hypertable overview[68]

[66] Cf. DB-Engines (w.y.)
[67] Cf. Hypertable Inc. (2015)
[68] Cf. Hypertable Inc. (w.y.)

The Hyperspace service shown on the right, the equivalent of Google's Chubby service. A lock service, which provides a file system for metadata. As Google states it is "the root of all distributed data structures".[69] Every newly instantiated server needs to first contact Hyperspace before beginning to work as a database server. The master is in charge of creating tables, deleting tables and all similar meta operations. Furthermore it detects failures of the RangeServers and assigns new ranges if needed. The just mentioned RangeServers are in charge of read and write operations. They also hold data in form of freely movable ranges, which as stated above are arranged by the Master. The File System Broker (FS Broker) is the process in which normalizes the underlying file system. Hypertable can rely on multiple file systems, for example file systems referred to in Figure 9. A FS Broker therefore may be seen as a file system translator which between the actual file system and the actual database. To ensure an easier access for developers Hypertable also provides an application interface called ThriftBroker.

Overall HyperTable is a free very transparent wide column database solution, which does not limit developers. The question why it is not as popular as other wide column databases remains unanswered. It might be due to a later release date and the rather small publishing company.

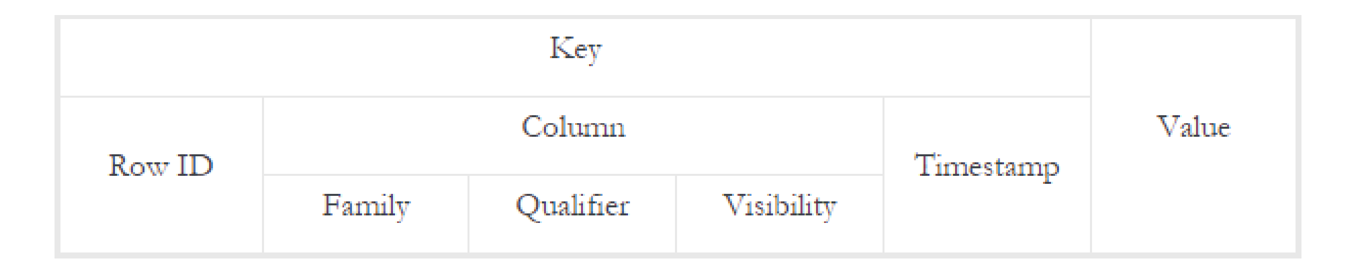
3.1.4 Accumulo

Accumulo is a database system that was submitted to the Apache Software Foundation in 2011, after being in development by the National Security Agency since 2008.[70] The main difference between Accumulo and other database systems is its fine-grained access control, down to cell-level (or better row-level) security. It is based on the Google BigTable and operates on the Hadoop Distributed File System (HDFS). [71]

[69] Cf. Hypertable Inc. (w.y.)
[70] Hoover, J. N. (2011)
[71] Apache Accumulo (2014)

Data Model

Tables in Accumulo consist of sorted key-value pairs. The key is comprised with the following elements:



Key					Value
Row ID	Column			Timestamp	
	Family	Qualifier	Visibility		

Figure 10: Key elements[72]

These key-value pairs are sorted by element and its lexicographical place in ascending order, whereas timestamps are sorted in descending order. This mechanism ensures that later versions of the same key appear first on a sequential selection of data.[73]

Architecture

Accumulo is a distributed database system that consist of different components:

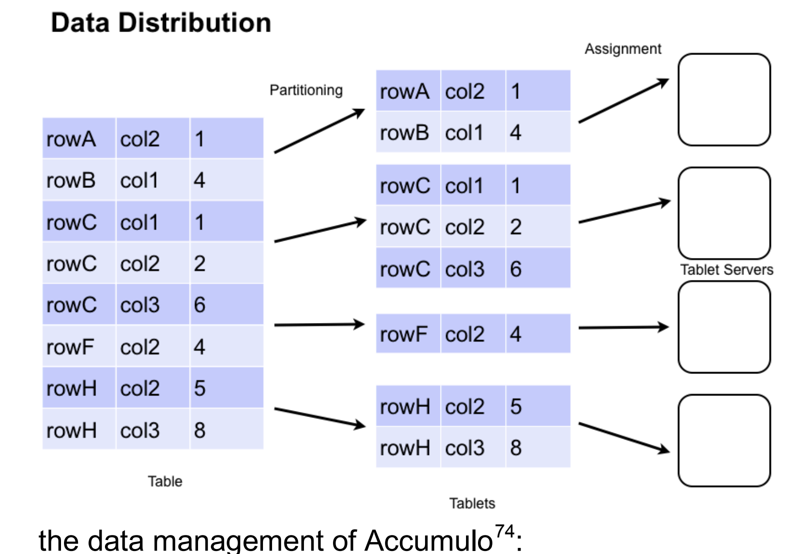
- Tablet Server
  - Manages partitions of Tables. It is responsible for the following tasks:
    - Managing writes from clients
    - Persisting writes to a so-called write-ahead log
    - Sorting new key-value pairs in memory
    - Flushing sorted key-value pairs into the HDFS periodically
    - Responding to read-requests from clients
    - Forming views of all keys and values of all files that have been created and sorted in memory
    - Recover tablets from failed server, using the write-ahead log
- Garbage Collector
  - Acumulo constantly shares files stored in the HDFS, files that are no longer needed are identified and deleted by the Garbage Collector

[72] Contained in: Apache Accumulo (2014)
[73] Apache Accumulo (2014)

- Master
  - It is possible to run several Masters, in which case they decide among themselves, which one becomes the single Master. The redundant Masters will act as backups in case of failure. The Master has the following objectives:
    - Detecting and responding to Tablet failure
    - Load-balancing between Tablets
    - Handling table creation, alteration and deletion request made by Clients
    - Coordinating startup, shutdown and recovery of failed Tablet Servers
- Client
  - Consists of an included client library that manages communication between applications (clients) and Accumolo

Data Management

Accumulo organizes its data in tables and splits those table additionally into tablets. A tablet is simply a smaller unit with several rows of any bigger table. The tables are portioned by row boundaries to ensure, that several columns of a given row are located within the same tablet. The Master only assigns tablets to one given Tablet Server at a time, which renders mechanisms for synchronization or locking of datasets unnecessary. The following picture clarifies



the data management of Accumulo[74]:

Figure 11: Data management of Accumulo[75]

[74] Apache Accumulo (2014)

[75] Contained in: Apache Accumulo (2014)

Notable Features

Notable features of Accumulo include:

- Tablet service
  - All write requests are also written to the so-called Write-Ahead Log. This Log is inserted into a MemTable. If the MemTa ble reaches a certain size its already sorted key-value pairs are written into a special file named Indexed Sequential Access Method file (ISAM). A note of these operations is made in the Write-Ahead Log. A request to read data is first executed on the MemTable to find the relevant indexes associated with the ISAM file to link them to the relevant values. The key-value pair is then returned to the client from the MemTable and a set of ISAM files.
- Compactions
  - To manage the growing number of files in a given table, the Tablet Server compacts several ISAM files into one. Previous files will be deleted by the garbage collector.
- Splitting
  - Based on the size of a table it will eventually be split into tablets. The new tablet is likely to be migrated to a different Tablet Server in order to keep the load of one given server to a minimum.
- Fault tolerance
  - If a given Tablet Server fails all of its write operations are extracted from the Write-Ahead Log and reapplied to another available server.
- Security
  - Each key-value pair in Accumulo has its own security label, stored in the column visibility element of a row. This determines if the user requesting the data within the row has sufficient authorization.[76]

Accumulo and HBase

Both Acculmulo and HBase are Apache licensed and run on the Apache Hadoop file system. Forums on the internet claim that both database systems are mainly the same, differing only in details. To further underline this argument the following table links Accumulo terminology to the corresponding one in HBase[77]:

[76] Apache Accumulo (2014)
[77] Apache Accumulo (2014)

Accumulo	HBase
Tablet	Region
Tablet Server	Region Server
Write-Ahead Log	HLog
Compaction	Flush

Table 6: Accumulo and HBase in comparison[78]

3.1.5 Sqrrl

Sqrrl was initially released in 2012 by Sqrrl Data, Inc., by a team which consists of several former National Security Agency alumni. Like others, the commercial database is based on Hadoop. Furthermore it is incredibly close to Accumulo. According to the Sqrrl homepage and their slogan “Securely Explore Your Data” it sets its emphasis on the security aspect of Big Data, especially since it also includes the cell-level security of Accumulo.[79] Moreover it implements Data-Centric Security, Role-Based Access Control and Attribute-based Access Control.[80] Figure 12 visualizes the architecture of Sqrrl. It is mainly based on Hadoop. Sqrrl extends this basis with the companies’ own Security, which was already describes above and is considerably similar to Accumulo. A big part of Sqrrl strategy is to deliver easy to understand solutions to make more of one’s data. This analytics philosophy is reflected by the available data models, interfaces, query languages and the data processing. It is unclear if these solutions are of a Sqrrl-native nature or just independent applications build on top.

[78] Contained in: Cloudera (2012)
[79] Cf. Sqrrl Data Inc. (w.y.)
[80] Cf. DB-Engines w.y.b)

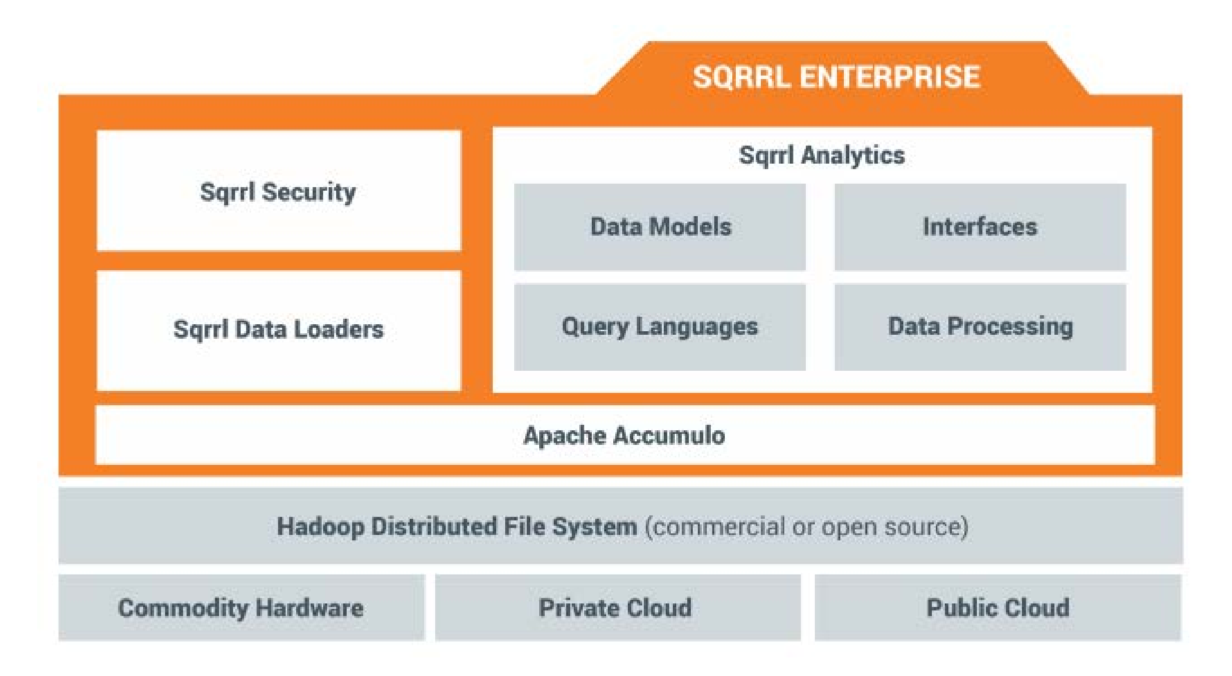


Figure 12: Sqrrl Enterprise Architecture[81]

The implementation language is Java, the supported operating system is Linux. It supports various programming languages and access methods. Also it implements MapReduce capabilities. On the corresponding data sheet Sqrrl promises simple Big Data analysis with a special focus on cyber security, healthcare analytics or data-centric security. The extraordinary strong focus on security and the apparently elementary administration of the user interface seem to be the only unique selling point. Especially since this commercial product comes with secrets and even after inquiries remains a “black box”. This remains a very strong point, particularly since there are very good transparent tools, which additionally are free of charge.

3.2 Comparison of database products on the basis of a list of criteria

In order to evaluate the databases systems introduced in chapter 3.1. Common criteria are chosen to establish a basic comparability between the products and finally decide which system to implement as a prototype for further examination. The list of criteria is mainly based on the official documentation of each database to deliver the most exact results possible. The list is separated into four major fields: general information, design and features, performance and administration. Note that the list of criteria is established before the examination of available features in the documentations, therefore it consist of general requirements of a database system. Consequently this means, that several fields could not be filled out, because of missing information. These fields are marked red to highlight the absence of information.

[81] Sensmeier (w.y.), p. 2

General criteria

- This section of the list of criteria contains basic and general information of each database, such as the name, license, a short description and most importantly the overall popularity of the product.

Annotation to determine the popularity of a given database

An important factor to consider, before making the decision which database system to implement, is the current popularity or use in businesses. More successful systems tend to offer more resources in form of for example available trained personnel or documentation and updates. In order to quantify the popularity of each database introduced, its db-engines.com ranking is presented in the catalog of criteria. It is nearly impossible to determine the exact number of installed bases, the databases are ranked on said score instead.

db-engines.com determines the popularity score on the basis of the following factors:

- Number of mentions of the system on websites
    - Based on results by Google and Bing
- General interest in the system
    - Based on Google Trends
- Frequency of technical discussions about the system
    - Based on the number of related topics and interested users on websites like Stack Overflow and DBA Stack Exchange
- Number of job offers, where the given system is mentioned
    - Based on job search engines
- Number of profiles in professional networks like LinkedIn
- Relevance in social networks
    - Based on data from websites like Twitter

Overall the criteria used offer a well-rounded view on the popularity of a system and therefore possibly the use in organizations.

Design and feature criteria

- This section of the list contains information about the implementation language, supported operating systems and programming languages, the database model (for those databases that are not only wide-column stores natively) and important features of wide-column stores like the availability of multiplexing (clustering).

Performance criteria

- The basic idea of the performance section is to compare the number of read or write operations per time unit for each system. While a good sounding concept in theory, it proved impossible to find comparable pre-established data on the internet to draw meaningful conclusions from.

Administration criteria

- This section of the list focuses on the administration of the database, with criteria such as the used integrity level, the security possibilities in form of user account control or available backup features.

Decision for a suitable prototype

The decision on which prototype to implement concluded in favor of Apache Cassandra for the following reasons:

- The overall availability of online documentation of Cassandra seems the best when compared to the other examined products, therefore it should be easiest to implement Cassandra for organizations who wish to explore wide-column store databases
- Its high score according to the db-engines.com rating, which underlines the statement made above and suggest a useful production system
- Cassandra claims to be the easiest system to set up a multi-node cluster. This is especially important for wide-column stores, because their design-philosophy is based on the premise of a distributed system
- Cassandra is the system the most information could be gathered on, all other databases have gaps in available descriptions on the internet, which makes the idea of building a production system around it a somewhat tedious work with a lot of trial and error implementations, which are unnecessary wasted resources

3.3 Implementation of prototype

One of the main mission of this paper is to test the implementation of a wide-column NoSQL database, namely, as discussed in the previous chapter, Apache Cassandra, and compare it to a traditional relational database, namely MySQL. Furthermore some testing will be conducted, to examine the basic performance promises by the developer.

Both prototypes will be set up on identical environments in order to achieve comparable results. Because of limited time and hardware possibilities the systems will run virtualized in VMware Workstation 10.0.3. The host and prototype systems have the following specifications:

	Host system	Prototype system
Operating System	Windows 8.1 Enterprise (64-bit)	Ubuntu Desktop 14.4.1 (64-bit)
CPU speed	Core i5 2520M @ 2,5 GHz	1 CPU with 2 cores
RAM	16 GB	2 GB
Hard drive	SSD 250 GB	20 GB

Table 7: Host and prototype system specifications

Ubuntu is chosen as the suitable option for setting up the test environment. It is well supported with third-party software, stable and most importantly is based on a Linux kernel. The Linux installation is the only officially supported version of Cassandra. Additionally it is assumed that companies would set up a Cassandra system on a Linux or Unix system themselves. Therefore the Linux installation will serve as a proper test for future production systems.

3.3.1 Cassandra

Cassandra was designed to be run as a distributed system of multiple Cassandra nodes. To test possible differences between a single- and multi-node systems it becomes apparent that two prototypes of Cassandra have to be set up. In detail a single-node and a three node system. The following instructions assume that the operating system has been set up functionally.

Pre-configuration of the system

In order to install Cassandra properly, several prerequisites have to be met. All commands are executed in the Terminal of Ubuntu.

1. **Install Java:** The current JDK is installed. It is important to note that Cassandra works best with the Linux JDK offered by Oracle. The OpenJDK is known to have issues with Cassandra. The following commands are executed

```
sudo add-apt-repository ppa:webupd8team/java
sudo apt-get update
sudo apt-get install oracle-java7-installer
```

2. **Set up the Java environment variable:** The environment file is edited by executing the following

```
sudo nano /etc/environment
add the line:
JAVA_HOME=/usr/lib/jvm/java-7-oracle
```

The system is restarted

```
sudo shutdown -r now
```

3. **Set up directory:** A directory for later install and easy operation of Cassandra is set up

```
sudo mkdir /root/home/cassandra
```

4. **Download Cassandra:** The Cassandra tarball files are downloaded

```
wget
http://planetcassandra.org/cassandra/?dlink=http://downloads.datastax.c
om/community/dsc-cassandra-2.0.11-bin.tar.gz
```

Install Cassandra

5. **Install Cassandra:** The tarball version of Cassandra is simply installed by extracting the .tar.gz file to the desired directory

```
tar -xvzf dsc-cassandra-2.0.11-bin.tar.gz
mv apache-casandra-2.0.3/* /root/home/cassandra
```

6. **Set up data directories:** Cassandra sets up several directories to store runtime relevant data in. The location of these directories is determined by entries in the central configuration file of Cassandra, the cassandra.yaml file, located in cassandra/conf. This file is edited by

```
nano cassandra.yaml
```

The following entries are changed to

```
data_files_directory: - /root/cassandra/data

commitlog_directory: /root/Cassandra/commitlog

saved_caches_directory: /root/cassandra/saved_caches
```

7. **Start Cassandra:** Cassandra is started by executing the cassandra file in the cassandra/bin directory

```
./cassandra
```

Set up the three node cluster

1. **Set up cluster:** Cassandra clusters are set up by installing Cassandra on every individual node and furthermore set parameters in the cassandra.yaml file to point to other nodes. There are basically two types of nodes, normal nodes and seeds. Seeds represent relatively stable nodes, which serve as controllers for other nodes. They manage the data flow between single instances. The following steps are executed

```
nano Cassandra.yaml
```

The three node system will consist of two normal nodes and one seed. The seed is set up first. It is curtail to set up a proper network with ideally static IP-addresses. The test system has the following setup:

	IP-address
Node 0 (seed)	192.168.123.125
Node 1	192.168.123.126
Node 2	192.168.123.127

Table 8: Setup of test system

The following entries in the cassandra.yaml file are changed

Parameter	Node 0	Node 1	Node 2
seeds:"…"	"192.168.123.125"	"192.168.123.125"	"192.168.123.125"
listen_address	192.168.123.125	192.168.123.126	192.168.123.127
rpc_address	0.0.0.0	0.0.0.0	0.0.0.0
snitch	RackInferringSnitch	RackInferringSnitch	RackInferringSnitch

Table 9: Entries in cassandra.yaml file

The seed entry always points to the IP-address of the node that is determined as the seed. The listen_address determines the IP that other nodes have to contact to communicate with this node, ergo its own IP. The rpc_address serves as the listen address for remote procedure calls. Setting it to 0.0.0.0 configures it to listen to all available interfaces. The snitch is a protocol to locate and route requests. The RackInferringSnitch determines that all configured nodes are part of a single logical datacenter.

Annotation to available documentation and versions

The official Apache Cassandra documentation (http://wiki.apache.org/cassandra/FrontPage) claims that Cassandra is accessible to set up and operate. However the content of this official documentation is inadequate at best. Several topics required for a proper set up are simply missing or just superficially described, for example the instructions on how to set up a cluster are mentioning important tags, but do never explain how or where to change these parameters.

Another possible source of information is the planetcassandra.org website, mainly operated the company Datastax. Datastax is contributing heavily to the Cassandra project, up to the point where they offer their own version of Cassandra, which contains additional tools for a more accessible setup. The main issue here is that one has to register in the Datastax community to access more detailed training on how to set up components, for example clusters. The company offers an open source version, available to everyone, and sells a business solution. Only the business solution contains all the documentation necessary to implement a

useful production environment with Cassandra. Therefore it is very tedious work to gather all required information concerning Cassandra and is linked to a lot of trial and error runs before a functionally system is set up. The official and publicly accessible documentation is insufficient.

For convenience this paper uses the open source version of Cassandra offered in the planetcassandra.org website by Datastax.

3.3.2 MySQL relational database

To compare the wide column database with a relational database a MySQL database is installed for testing. MySQL is a popular open source relational database management system and used by many companies like Facebook and Google.[82] XAMPP is used as a development environment for MySQL and is also open source.[83] The scenario that is explained in the chapter "Data Feeders" is used for this relational database. A script written in Java is constantly generating data and inserts them through a JDBC database connection into the MySQL database. The relational database is as similar as it can get to the wide column database so it can be compared in terms of performance, scalability and flexibility.

In order to set up the MySQL database on the Ubuntu test system, XAMPP is installed first:

The latest version of XAMPP is downloaded

```
Wget
http://downloads.sourceforge.net/project/xampp/XAMPP%20Linux/5.5.19/xam
pp-linux-x64-5.5.19-0-installer.run
```

XAMPP is installed by executing

```
chmod 755 xampp-linux-VERSION-installer.run

sudo ./ xampp-linux-VERSION-installer.run
```

XAMPP can be run by executing

```
sudo /opt/lampp/lamp start
```

After starting the server, the interface is available by http://localhost in the browser, the MySQL database can be accessed by using the MyPHPadmin link in the appearing menu.

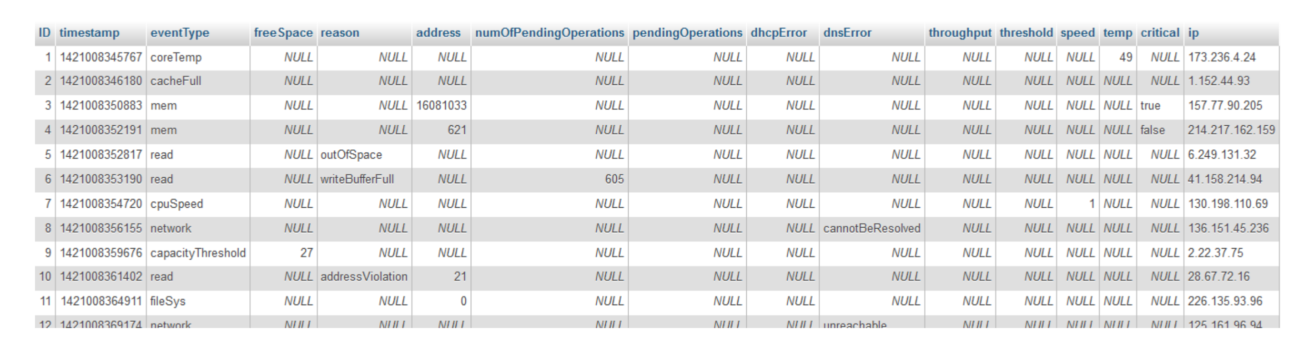
A new database is created with the name "db_projekt" and a new relation "events" was created with different attributes. The schema has to be predefined with the Data Definition Language so our columns are as follows:

[82] MySQL (w. y.a)
[83] Apache Friends (w. y.a)

Name	Data type
ID	bigint auto_increment primary key
Timestamp	Bigint
eventType	Barchar
freeSpace	Int
Reason	Barchar
Address	Int
numOfPendingOperations	Int
pendingOperations	Int
dhcpError	Varchar
dnsError	Varchar
Throughput	Int
Threshold	Int
Speed	Int
Temp	Int
Critical	Varchar
Ip	Varchar

Table 10: SQL attributes

The timestamp is quite high so the data type has to be in this case a bigint.



ID	timestamp	eventType	freeSpace	reason	address	numOfPendingOperations	pendingOperations	dhcpError	dnsError	throughput	threshold	speed	temp	critical	ip
1	1421008345767	coreTemp	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	49	NULL	173.236.4.24
2	1421008346180	cacheFull	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	1.152.44.93
3	1421008350883	mem	NULL	NULL	16081033	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	true	157.77.90.205
4	1421008352191	mem	NULL	NULL	621	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	false	214.217.162.159
5	1421008352817	read	NULL	outOfSpace	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	6.249.131.32
6	1421008353190	read	NULL	writeBufferFull	NULL	605	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	41.158.214.94
7	1421008354720	cpuSpeed	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	1	NULL	NULL	130.198.110.69
8	1421008356155	network	NULL	NULL	NULL	NULL	NULL	NULL	cannotBeResolved	NULL	NULL	NULL	NULL	NULL	136.151.45.236
9	1421008359676	capacityThreshold	27	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	2.22.37.75
10	1421008361402	read	NULL	addressViolation	21	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	28.67.72.16
11	1421008364911	fileSys	NULL	NULL	0	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	226.135.93.96
12	1421008369174	network	NULL	NULL	NULL	NULL	NULL	NULL	unreachable	NULL	NULL	NULL	NULL	NULL	125.161.96.94

Figure 13: Screenshot of MySQL

The data feeder is responsible for creating new tuples and it inserts them with the Data Manipulation Language into the relational database.

3.4 Testing of prototype systems

The testing is designed to examine the performance of wide-column NoSQL database systems with the example of Apache Cassandra in comparison to traditional relational database system on the example of MySQL. The most probable case of deployment of a wide-column NoSQL system being when associated Big Data is not likely to be verified on a virtual test environment such as the one implemented in previous chapters by this paper. Therefore this test focuses on a basic and simple command execution and performance comparison.

The business case these tests are related to is the logging of system events, which should be written to a database for later processing.

3.4.1 Test design

The tests focus on two operation principles of databases, the read and write commands, more detailed the execution time of the INSERT and SELECT of rows. For these purposes a Java application, which sets up the data model inside the database, inserts and selects data based on system events is developed. This application will be discussed more detailed in a later chapter.

The times measured is the actual overall execution time of said Java application and the accumulated time of executing a CQL/SQL statement by the database. Every test is run on the Cassandra single-node system, the Cassandra multi-node system and the MySQL system several times to finally raise an averaged result set.

Write testing by INSERT

	Description
Write test 1	Write system events, which are randomly generated by the Java feeder application, to the Cassandra single-node system. The testing interval is 100, 10.000, 100.000 and 1.000.000 [rows to write per execution].
Write test 2	Write system events, which are randomly generated by the Java feeder application, to the Cassandra multi-node system. The testing interval is 100, 10.000, 100.000 and 1.000.000 [rows to write per execution].
Write test 3	Write system events, which are randomly generated by the Java feeder application, to the MySQL system. The testing interval is 100, 10.000, 100.000 and 1.000.000 [rows to write per execution].

Table 11: Write testing by INSERT

Read testing by SELECT

	Description
Read test 1	Read the events written to the Cassandra single-node system in the “write testing by INSERT”-tests by SELECT. The following statements are executed sequentially: 1. SELECT * FROM events; 2. SELECT * FROM events WHERE eventType='mem';
Read test 2	Read the events written to the Cassandra multi-node system in the “write testing by INSERT”-tests by SELECT. The following statements are executed sequentially: 3. SELECT * FROM events; 4. SELECT * FROM events WHERE eventType='mem';
Read test 3	Read the events written to the MySQL system in the “write testing by INSERT”-tests by SELECT. The following statements are executed sequentially: 5. SELECT * FROM events; 6. SELECT * FROM events WHERE eventType='mem';

Table 12: Read testing by SELECT

3.4.2 Data feeders

The scenario deals with an Event-Logger, which is placed in a fictitious data center to monitor all kind of equipment like servers. If the application identifies runtime issues, by comparing different values to a pre-set benchmark, it saves an event report to the database. Each report in the database contains an identification number (ID), a timestamp of when the device spotted the issue, the device’s Internet-Protocol (IP) address, the event type and optional extra metrics. The below table shows possible event types and their extra metrics:

Event Type	Description	Extra Metrics
capacityThreshold	Disk space almost full	- Percentage of free space
Write	Value could not be written do disk	- Reason - Address - Number of pending operations
Read	Value could not be read from disk	- Reason - Address - Number of pending operations
Mem	Memory error	- Address - Critical (yes/no)
coreTemp	CPU core temperature high	- Temperature in °C
fileSys	General file system error	- Address
Network	Network error	- dhcpError (with detailed reason) - dnsError (with detailed reason) - threshold (with percentage of network usage)

Kernel	Kernel error	- Address
cpuSpeed	CPU speed to low	- Speed in mHz
cacheFull	Cache is full	

Table 13: Event type

The scenario described has been selected, because it features continues input of mass information. The data is also not perfectly structured and it is highly likely that other "columns" may be added during runtime. This scenario poses some major difficulties to traditional database management systems and is supposed to show whether NoSQL-databases can better deal with this common scenario.

Important features of this feeder-application include:

- Randomized generation of system events; each write to the database is constructed of a pool of available parameters in order to simulate dynamic data input
- Different parts of the application contain a timer, which is later output to a .txt file. This ensures the timing with minimal latencies, because of its proximity to the code. Times measured include for example the overall execution time of the application or the accumulated time spent waiting for database responses

3.4.3 Execution of tests

The tests are executed by running the corresponding Java application on every test system, assuming Cassandra and MySQL have been set up properly and are running.

Cassandra single-node

- Execute the FeederCreator.jar to set up the data model in the database by

```
java –jar /location/FeederCreator.jar 1
```

- Run the write test by executing the FeederCassandra.jar sequentially

`java –jar /location/FeederCassandra.jar 100` (10.000, 100.000, 1.000.000)

- Run the read test by executing the ExecutionTimerCassandra.jar with the following parameters

`Java –jar /location/ ExecutionTimerCassandra.jar "SELECT * FROM events;"` ("SELECT * FROM events WHERE eventType='mem';)

Cassandra multi-node

- Execute the FeederCreator.jar to set up the data model in the database by

```
java –jar /location/FeederCreator.jar 1
```

- Run the write test by executing the FeederCassandra.jar sequentially

`java –jar /location/FeederCassandra.jar 100` (10.000, 100.000, 1.000.000)

- Run the read test by executing the ExecutionTimerCassandra.jar with the following parameters

`Java –jar /location/ ExecutionTimerCassandra.jar “SELECT * FROM events;”` (“SELECT * FROM events WHERE eventType='mem';)

MySQL database

- Run the write test by executing the FeederMySQL.jar sequentially

`java –jar /location/FeederMySQL.jar 100` (10.000, 100.000, 1.000.000)

- Run the read test by executing the ExecutionTimerMySQL.jar with the following parameters

`Java –jar /location/ ExecutionTimerMySQL.jar “SELECT * FROM events;”` (“SELECT * FROM events WHERE eventType='mem';)

All results generated by the testing application is output into the runMetrics.txt file.

3.5 Results of testing

For write tests two timings were measured. ‘TotalTimeSpent’ is the entire runtime of the test application. ‘TimeSpentWaiting’ is the accumulated time spent waiting for a response from the database server during an execution, taking respect of the fact that not the entire runtime is spent for the database.

Figure 14 shows the TotalTimeSpent results for the write tests. Detailed results can be found in Table 14. As the number of inserted rows per test run grows, so does the total execution time. The fastest prototype is the Cassandra single node, followed by the MySQL database. The Cassandra multi node prototype is clearly slower than the other solutions. It seems quite surprising that MySQL lies in between the two Cassandra prototypes and that the multi node’s execution time is three times the single node’s. As one recollects the fact that due to resource scarcity the multi node has been set up as three virtual machines on one single host system with the feeder running on one of these systems, an obvious explanation comes to mind.

The host's system power is split over three systems, which means that each of them cannot use the full host's power. Clearly, the system will have less resources than the single node prototype. So when multi node configuration is used either the database engine or the feeder application or both seem to be slower than in single node configuration.

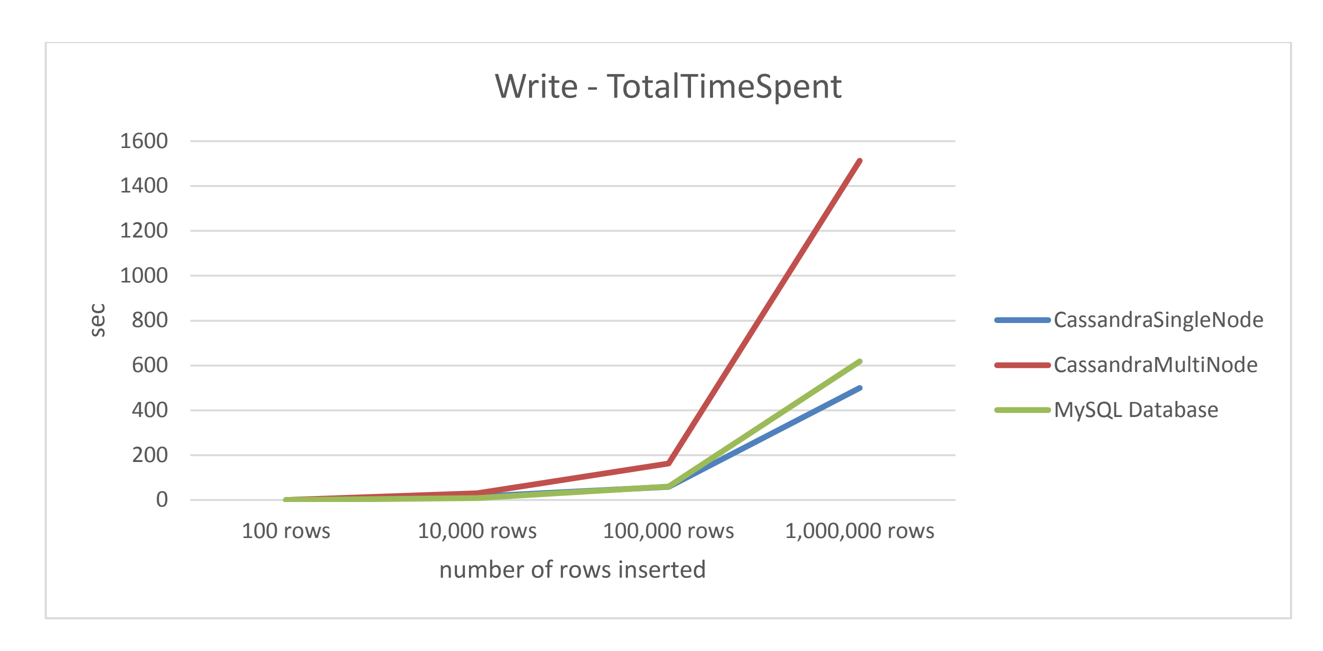


Figure 14: Total Time Spent for Write Tests

Write: TotalTimeSpent [sec]	100 rows	10,000 rows	100,000 rows	1,000,000 rows
CassandraSingleNode	0,482581102	16,0358748	59,131706932	499,171362974
CassandraMultiNode	0,46103862	30,62567455	162,606194683	1.513,54875017
MySQL database	0,25076575	8,902223879	60,50125513	618,6001182

Table 14: Total Time Spent for Write Tests

The time spent waiting for a response from the database during write tests is illustrated in Figure 15, detailed results can be found in Table 15. As the number of inserted rows per execution grows, the time spent waiting for the database shrinks. For the Cassandra multi node prototype and the MySQL prototype the shrinking time spent waiting reaches a turning point at 100,000 rows per execution and starts growing again. On the whole, the Cassandra single node prototype again delivers best results, followed by MySQL in the middle and the Cassandra multi node as the worst. Again, at first glance the results are surprising. As these are accumulated waiting times, the times are not supposed to shrink while the number of iterations grows. 100 executions can be done more efficient than 10, but they can never be faster. Again, the solution of this paradox lies in system performance, more precisely the feeder.

Assume for a great number of iterations java's efficiency shrinks, as it is busy running the same code again and again. This would mean that the feeder can send less requests per second to the database. As only the waiting time is measured, the values shrink, because java gives the database more time to answer and thus there is less waiting time.

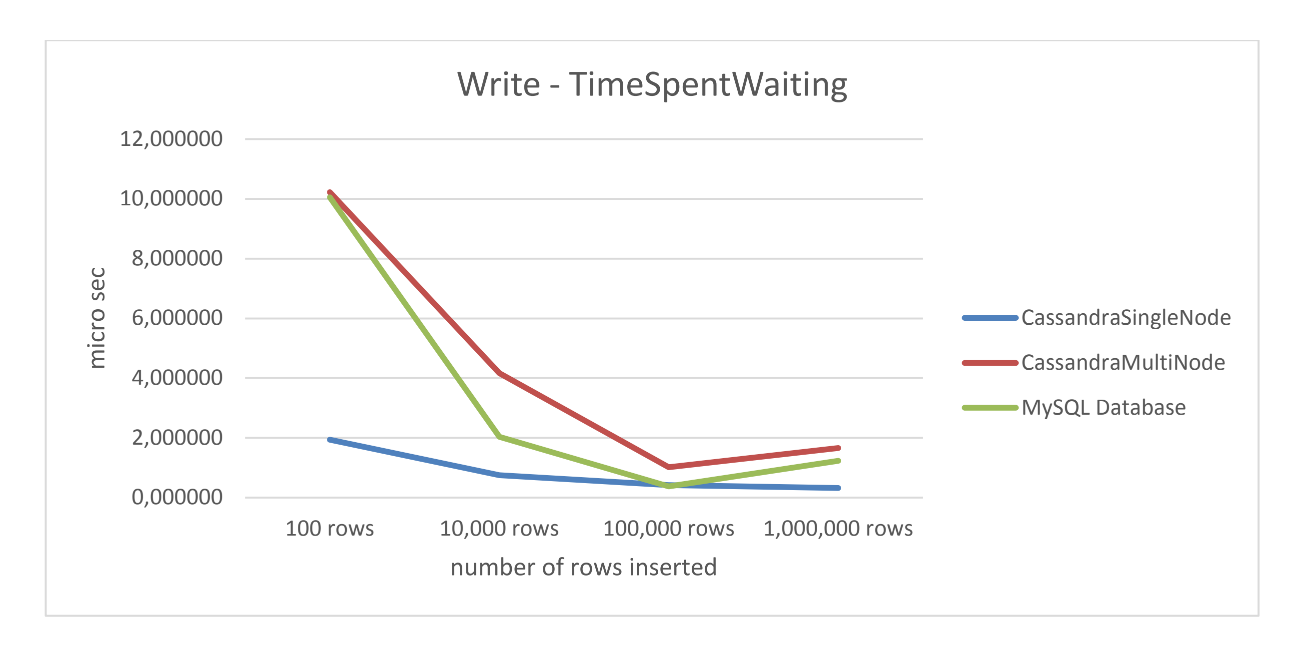


Figure 15: Time Spent Waiting for the Database during Write Tests

Write: TimeSpentWaiting [micro sec]	100 rows	10,000 rows	100,000 rows	1,000,000 rows
CassandraSingleNode	1,942136	0,752343	0,411938	0,317332
CassandraMultiNode	10,227307	4,165838	1,016842	1,663259
MySQL database	10,057036	2,034421	0,370646	1,230051

Table 15: Time Spent Waiting for the Database during Write Tests

In the test environment the Cassandra database is slower in multi node configuration because of the resource scarcity. For growing numbers of iterations per execution the total runtime grows and the waiting time shrinks. The latter is caused by a slower feeder application which gives the database more time to respond before waiting time is recorded. The slowest solution, the Cassandra multi node prototype, does not deliver the smallest waiting time, as not only the feeder but also the database itself slows down.

A comparison of TimeSpentWaiting and TotalTimeSpent can give some final insights. Again, results are illustrated in Figure 16 and detailed numbers can be found in Table 16. As the number of inserted rows per execution grows, the portion of waiting time per total runtime shrinks or the efficiency grows.

The most efficient solution is the Casandra single node prototype, followed by the Cassandra multi node prototype and MySQL. Notice that in terms of efficiency the Cassandra prototypes rank 1 and 2 and MySQL ranks 3. Earlier, MySQL has always been in between the two Cassandra prototypes.

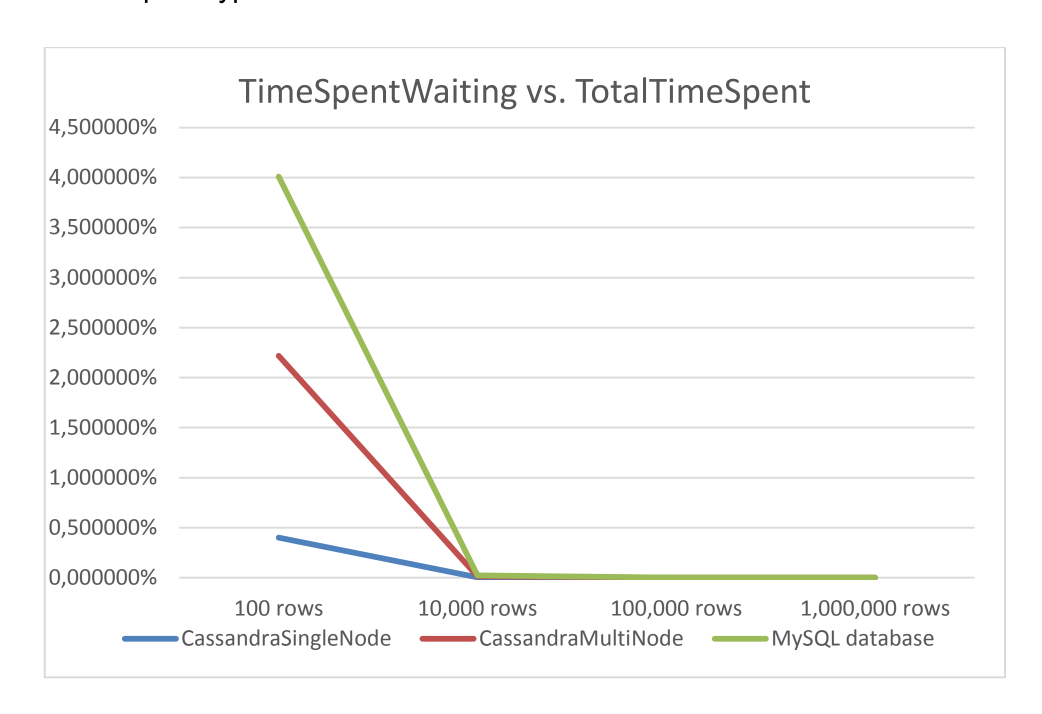


Figure 16: TimeSpentWaiting vs. TotalTimeSpent

Write: WaitingVsTotal	100 rows	10,000 rows	100,000 rows	1,000,000 rows
CassandraSingleNode	0,402448%	0,004692%	0,000697%	0,000064%
CassandraMultiNode	2,218319%	0,013602%	0,000625%	0,000110%
MySQL database	4,010530%	0,022853%	0,000613%	0,000199%

Table 16: TimeSpentWaiting vs. TotalTimeSpent

Finally, the results of the read tests can be seen in Table 17. As these test applications consist only of one command, total runtime equals time spent waiting. For selecting rows of a table when applying a filter ('SQL 2'), the Cassandra single node prototype was the fastest, followed by MySQL and the multi node solution. When requesting an ordered output, only MySQL responded. Ordered output is only possible in Cassandra when the ordered column is the primary key – this was not the case in the test scenario.

Read: TimeSpentWaiting [micro sec]	SQL 2	SQL 2 + ORDER BY
CassandraSingleNode	1501,485301	N/A
CassandraMultiNode	10485,16365	N/A
MySQL database	2380,667349	1411,508593

Table 17: Time Spent Waiting for Reads

3.5.1 Comparison to the End Point Benchmark

End Point, a database consulting company, conducted tests on the top NoSQL database engines. The tests were run using a variety of workloads on Amazon Web Services EC2 instances. So in contrast to the test environment used in this paper End Point did not face the problem of resource scarcity. One of their test results was that a single Cassandra instance can handle about 7,000 writes per second – in the above tests the prototypes reached around 500 writes per second. As expected, in End Point's tests throughput increases as the number of nodes increases. Unfortunately, End Point does not compare Cassandra to MySQL. But throughout all disciplines, Cassandra delivers the best performance among NoSQL databases.[84]

3.5.2 Review of Implementation and testing

The conducted tests delivered interesting and partly unexpected data. For all unexpected results there are possible solutions, but the discussed resource scarcity poses a threat to the reliability of the results. The multi node solution could not be tested adequately, as all nodes ran as virtual machines on one host system, hence dividing resources by three. Throughout all tests the Cassandra single node prototype was the most powerful solution and in most tests MySQL ranked second.

Comparing read results posed special problems as CQL and SQL are not as close as one could assume from reading the Cassandra documentation. CQL only provides basic queries excluding aggregate functions, ordering by any column other than the primary key, or grouping. Even filtering by rows using the WHERE clause is only possible in CQL if an index has been manually created on the mentioned column.

The implementation showed that testing big data environments is not possible on small or weak systems. Furthermore it showed that for write requests Cassandra is more efficient than MySQL but for read requests MySQL is faster. In addition, it has been noticed that SQL cannot easily be turned into CQL, the underlying structure is completely different. Rich queries like in SQL are not possible in CQL which forces the database administrator to save the data in a way which supports the planned read queries.

[84] Planet Cassandra (2015d)

4 Conclusion

Deep insight into the top five wide column databases has been provided. Also a list of criteria has been developed and extensively discussed. Based on the list of criteria a wide column system, Apache Cassandra has been chosen for implementation and testing. Three prototypes – MySQL, Cassandra single node and Cassandra multi node – have been implemented. In order to do so, the project team had to gain deep knowledge on the Cassandra system and how it works. A test scenario has been developed and a random generator to feed the prototypes with data has been implemented. Six test types were executed on each of the prototypes and the results have been presented and discussed. In the end, a comparison to the End Point Benchmark took place. In conclusion, all objectives have been met.

One key finding is that there actually is no wide variety of different wide column databases. There indeed are numerous products, but most of them have been inspired by each other or even base on each other. Accumulo is a famous core for many products, Hadoop is the file system used by almost every product. Apache Cassandra has got its own file system and core, but it is based on Google's BigTable.

Anyways, Cassandra has got a quite impressive architecture. In contrast to RDBMS Cassandra provides native cluster support and real fault tolerance without a single point of failure. A Cassandra cluster can be seen as a swarm of machines: every node can be contacted by a user and will answer for the entire cluster. A tunable replication factor determines on how many different nodes a row will be stored, tunable consistency determines how many nodes have to send a positive answer before the entire request is considered positive. For replication Cassandra is also aware of a node's location, which makes it very easy to implement multiple data centers. Procedures like mirroring and load balancing can be configured on the fly and Cassandra manages backup and restore on its own. The engine is designed to run on cheap hardware and is highly scalable. In summary, Cassandra fits the requirements of the 21st century with big topics like mass data and decentralization.

On the other hand, Cassandra is not as flexible as expected when it comes to data structure and scheme. NoSQL databases promise that a user can "throw" all kind of data into the database and some columns may exist in one row but not in another. In fact, Cassandra's CQL has got similar requirements as SQL: inserting data is only possible if the respective column has been set up in a schema. This column then exists in every row of a table with null values representing the case that "a column does not exist in a row". Even data types need to be defined and Cassandra does not even accept a string containing a number for entering into am integer column. But then CQL does not provide SQL-like query features.

Aggregate functions and grouping are not possible, filtering is only possible for indexed columns and sorting is only possible for the primary key. The user is forced to save data in a manner that supports the planned read queries.

The tests unfortunately did not deliver one hundred percent trustworthy results. The limited resources of the testing environment caused that the prototypes stayed far below their capabilities. Also in multi node setup the virtual machines had to fight for resources with each other. However, all unwanted effects could be explained and a good efficiency evaluation could be done. To support the prototype evaluation the Big Point Benchmark has been discussed, which stated that Cassandra is the best performing wide column database.

This project paper met its objective to bring some light into the mystery of big data. Interesting insights were given and many findings could be achieved. Clearly it can be said that the shift from SQL to NoSQL is not easy for a database administrator and that it also requires a shift in thinking.

Publication bibliography

Home | Hypertable - Big Data. Big Performance. Available online at http://hypertable.com/home/, checked on 1/13/2015.

NoSQL Databases Defined & Explained. Available online at http://planetcassandra.org/what-is-nosql/, checked on 1/13/2015.

Scale Out with Hypertable. Available online at http://www.linux-mag.com/id/6645/, checked on 1/13/2015.

Architecture | Hypertable - Big Data. Big Performance (2015). Available online at http://hypertable.com/documentation/architecture/, updated on 1/14/2015, checked on 1/14/2015.

Why Hypertable? | Hypertable - Big Data. Big Performance (2015). Available online at http://hypertable.com/why_hypertable/, updated on 1/14/2015, checked on 1/14/2015.

Abadi, Daniel (2012): The Design and Implementation of Modern Column-Oriented Database Systems. In *FNT in Databases* 5 (3), pp. 197–280. DOI: 10.1561/1900000024.

Abadi, Daniel J. (Ed.) (2007): Column Stores for Wide and Sparse Data. Available online at http://web.mit.edu/tibbetts/Public/CIDR_2007_Proceedings/papers/cidr07p33.pdf.

Anderson, Jesse (2013): Intro To MapReduce. Available online at https://www.youtube.com/watch?v=bcjSe0xCHbE, checked on 1/20/2015.

Apache Accumulo (2014): Apache Accumulo User Manual Version 1.5. Available online at https://accumulo.apache.org/1.5/accumulo_user_manual.html, updated on 3/17/2014, checked on 1/22/2015.

Apache Friends (w. y.a): XAMPP Installers and Downloads for Apache Friends. Available online at https://www.apachefriends.org/index.html, checked on 1/15/2015.

Apache HBase (2015b): Chapter 1. Architecture. Available online at http://hbase.apache.org/book/architecture.html#arch.overview, updated on 12/22/2014, checked on 1/12/2015.

Apache HBase (2015a): HBase – Apache HBase™ Home. Available online at http://hbase.apache.org/, updated on 12/22/2014, checked on 1/12/2015.

Bhattacharjee, A. (2014): NoSQL vs SQL – Which is a Better Option? Available online at https://www.udemy.com/blog/nosql-vs-sql-2/, checked on 1/19/2015.

Cho, Terry (2010): Apache Cassandra Quick Tour. Available online at http://javamaster.files.wordpress.com/2010/03/cassandra_data_model.png, updated on 3/22/2010, checked on 1/20/2015.

Cloudera, Inc. (2012): HBase and Accumulo | Washington DC Hadoop User Group. Available online at http://de.slideshare.net/cloudera/h-base-and-accumulo-todd-lipcom-jan-25-2012, checked on 1/22/2015.

Datastax (2015a): CQL for Cassandra 2.x. Available online at http://www.datastax.com/documentation/cql/3.1/pdf/cql31.pdf, checked on 1/19/2015.

Datastax (2015b): XMP structure: 1. Available online at http://www.datastax.com/documentation/cassandra/2.0/pdf/cassandra20.pdf, checked on 1/19/2015.

DB-Engines (w.y.a): DB-Engines Ranking - popularity ranking of wide column stores. Available online at http://db-engines.com/en/ranking/wide+column+store, checked on 1/22/2015.

DB-Engines (w.y.b): Sqrrl System Properties. Available online at http://db-engines.com/en/system/Sqrrl, checked on 12/14/2014.

DeRoos, D. (2012): What is Big Data and how does it fit into an Information Integration Strategy. In Information Integration & Governance Forum 2012. Phoenix: IBM Corporation. Available online at http://194.196.36.29/events/wwe/ca/canada.nsf/vLookupPDFs/kitchener_-_bigdata_dirk/$file/Kitchener%20-%20BigData_Dirk.pdf, checked on 1/20/2015.

Dumbill, Edd (2012): What is big data? - O'Reilly Radar. California: O'Reilly Media. Available online at http://radar.oreilly.com/2012/01/what-is-big-data.html, checked on 1/20/2015.

George, Lars (2011): HBase: the definitive guide: " O'Reilly Media, Inc.".

Google Cloud Platform (w.y.): mapreduce_mapshuffle.png (600×323). Available online at https://cloud.google.com/appengine/docs/python/images/mapreduce_mapshuffle.png, updated on 9/22/2014, checked on 1/20/2015.

Grehan, Rick (2015): Big data showdown: Cassandra vs. HBase. Available online at http://www.infoworld.com/article/2610656/database/big-data-showdown--cassandra-vs--hbase.html, updated on 1/12/2015, checked on 1/12/2015.

Hoover J. N. (2011): NSA Submits Open Source, Secure Database To Apache - InformationWeek. Available online at http://www.informationweek.com/applications/nsa-submits-open-source-secure-database-to-apache/d/d-id/1099972?, checked on 1/22/2015.

Hypertable Inc. (w.y.a): Architecture. Available online at http://hypertable.com/documentation/architecture/, checked on 1/22/2015.

Hypertable Inc. (w.y.b): Company. Available online at http://hypertable.com/company/, checked on 1/21/2015.

IBM (w.y.): Accelerated analytics - faster aggregations using the IBM DB2 for i encoded vector index (EVI) technology. Available online at http://www.ibm.com/developerworks/ibmi/library/i-accelerated-analytics-db2-evi-tech/, updated on 10/31/2013, checked on 1/17/2015.

IBM Corporation (2014): IBM - What is HBase? Available online at http://www-01.ibm.com/software/data/infosphere/hadoop/hbase/, updated on 1/12/2015, checked on 1/12/2015.

Jhilam, Ray (2011): What is a Columnar Database? Available online at https://www.youtube.com/watch?v=mRvkikVuojU, checked on 12/6/2014.

Kroenke, David M.; Auer, David J. (2013): Database concepts. 7th: Pearson Education.

Kuznetsov, S. D.; Poskonin, A. V. (2014): NoSQL data management systems. In *Program Comput Soft* 40 (6), pp. 323–332. DOI: 10.1134/S0361768814060152.

Linwood, Jeff; Minter, Dave (2010): Beginning Hibernate. 2nd ed. [New York]: Apress. Available online at https://books.google.de/books?id=CW8mLHrLWnUC&printsec=frontcover&dq=beginning+hibernate&hl=de&sa=X&ei=D6a-VK7_AcrzauHYgrAH&ved=0CCUQ6AEwAA, checked on 1/20/2015.

mongoDB (w.y.): NoSQL Databases Explained. Available online at http://www.mongodb.com/nosql-explained, updated on 12/7/2014, checked on 12/7/2014.

mongoDB (2015): NoSQL Databases Explained. Available online at http://www.mongodb.com/nosql-explained?_ga=1.220084890.2012626732.1421159723, updated on 1/13/2015, checked on 1/13/2015.

MySQL (w. y.a): MySQL :: Why MySQL? Available online at http://www.mysql.com/why-mysql/, checked on 1/15/2015.

Pethuru Raj, Deka Ganesh Chandra: Handbook of Research on Cloud Infrastructures for Big Data Analytics (978-1-4666-5964-6). Available online at http://books.google.de/books?id=m95GAwAAQBAJ&pg=PA225&dq=wide+column+database&hl=en&sa=X&ei=kzN7VJWHD-LXyQP5oIDAAQ&ved=0CC8Q6AEwAQ#v=onepage&q=wide%20column%20database&f=false, checked on 1/13/2015.

Planet Cassandra (2015c): Data Replication in NoSQL Databases | Planet Cassandra. Available online at http://planetcassandra.org/data-replication-in-nosql-databases-explained/, checked on 1/20/2015.

Planet Cassandra (2015a): NoSQL Databases Defined & Explained. Available online at http://planetcassandra.org/what-is-nosql/, checked on 1/13/2015.

Planet Cassandra (2015d): NoSQL Performance Benchmarks: Cassandra vs HBase vs MongoDB vs Redis vs MySQL | Planet Cassandra. Available online at http://planetcassandra.org/nosql-performance-benchmarks/, checked on 1/22/2015.

Planet Cassandra (2015b): What is Apache Cassandra? | Planet Cassandra. Available online at http://planetcassandra.org/what-is-apache-cassandra/, checked on 1/19/2015.

Raj, P. (2014): Handbook of Research on Cloud Infrastructures for Big Data Analytics: IGI Global. Available online at https://books.google.de/books?id=m95GAwAAQBAJ.

Sadalage, P. J.; Fowler, M. (2012): NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence: Pearson Education. Available online at https://books.google.de/books?id=AyY1a6-k3PIC.

Scofield, B. (2010): NoSQL @ CodeMash 2010. Available online at http://www.slideshare.net/bscofield/nosql-codemash-2010, checked on 1/13/2015.

Sematext (2011): HBase Backup Options | Sematext Blog on WordPress.com. Available online at http://blog.sematext.com/2011/03/11/hbase-backup-options/, checked on 1/15/2015.

Sensmeier, Lisa (w.y.): Sqrrl_Hortonworks_ReferenceArchitecture-10292013. Available online at http://hortonworks.com/wp-content/uploads/2012/08/Sqrrl_Hortonworks_ReferenceArchitecture-102920131.pdf, checked on 12/14/2014.

Sqrrl Data Inc. (w.y.): Sqrrl | Securely Explore Your Data. Available online at http://sqrrl.com/, checked on 12/8/2014.

TechAmerica Foundation (w. y.a): Demystifying Big Data. Available online at http://mim.umd.edu/wp-content/uploads/2012/10/TechAmericaBigDataReport-2.pdf, checked on 1/20/2015.

The Economist (w. y.a): 201009SRC696.gif (GIF-Grafik, 290 × 281 Pixel). Available online at http://media.economist.com/images/20100227/201009SRC696.gif, updated on 2/24/2010, checked on 1/22/2015.

The Economist (2010): Data, data everywhere. Available online at http://www.economist.com/node/15557443, updated on 1/22/2015, checked on 1/22/2015.

Toad World (w. y.a): Toad World. Available online at http://www.toadworld.com/products/toad-for-cloud-databases/w/wiki/322.cassandra-column-families/revision/2.aspx, checked on 1/12/2015.

Appendix

- List of criteria



Einsatzszenarien von MVC-Frameworks zur Entwicklung client-seitiger SPAs

Schriftliche Ausarbeitung
im Rahmen der Lehrveranstaltung „Integrationsseminar“
für das Kompetenzzentrum Open Source (KOS)

Vorgelegt von

Andre Koch
Erik Mohr
Patrick Geppert

am 26.01.2015

Fakultät Wirtschaft
Studiengang Wirtschaftsinformatik
WWI2012I

Contents

List of abbreviations

CSS	Cascading Style Sheets
HTML	Hypertext Markup Language
MVC	Model-View-Controller
UI	User-Interface

List of figures

List of tables

1 Introduction

For programmers who want to develop client-centred applications nowadays the market offers a large amount of JavaScript frameworks. A popular trend here is the compliance of the MVC (Model-View-Controller) concept. Within the last few years several JavaScript frameworks have come up for development, usually being Open Source which means they are free to use for the programmers and developers.

All these frameworks kind of serve the same purpose of simplifying the development of application, yet there are quite some defining differences between them.

This work shall give a general insight about the technologies of web applications and their development. Further on the goal is to give an overview of the most important JavaScript frameworks existing at the moment. The main characteristics will be pointed out and afterwards a comparison between the different options will classify advantages and disadvantages in order to find the optimal technology to be used.

Picking out AngularJS this work will also show the development and implementation of an application using this framework and its functionalities.

2 MVC Frameworks

Before taking a closer look into the different MVC frameworks the Model-View-Control concept itself needs to be explained.

2.1 Model-View-Controller

Many computer systems are built to get data from a data store and display it. The use then can view the data and also change data. After that the system needs to update and store the new data in the data store. However the user interface usually changes much more frequently than the data storage system and the question is how to design the functionality of a Web application so that the individual parts can be modified easily.

The solution to this problem is the Model-View-Controller pattern separating the modelling of the data, the presentation and the user driven actions into three classes.

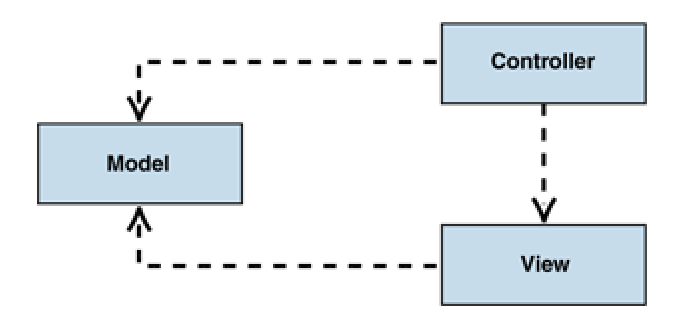


Figure 1: Structural Relationship[1]

The model represents the application core, e.g. a list of database records. It's the part that is responsible for the logic of the application data, managing the data and responding to requests for information or instructions to change.

The view is the part of the application that handles the display of the data. It's the actual visual element of the application also including things like control elements or animations and other effects.

[1] From: Microsoft Developer Network (2015)

The controller is all about the user interaction. It reads data from a view, controls user input and sends input data to the model. The controller reacts to mouse and keyboard commands and informs the model respectively the view about the input.

There are numerous advantages when implementing a solution with this pattern. A reduced amount of code directly connected to the user interface improves the testability of the application. A clean separation of the business logic (model) and the user interface (view) also reduces the risk of errors when migrating onto different devices since the user interface code is much more dependant in that regard. Further it is rather unlikely that one has both the skill to develop an appealing visual part and the complex business logic. Therefore it is reasonable to separate the development of these two parts. In conclusion it can be said that the MVC is a meaningful software design pattern for the development of Web applications using the division of UI logic from business logic.[2]

2.2 Comparison of JavaScript MVC frameworks

2.2.1 Introduction

In the last years JavaScript has been continuously the leader of client interactions on the web[3]. A main reason is the introduction of the smartphones and the difficult usability of Adobe Flash on these mobile devices. Therefore Apple Inc. restricted the use of Adobe Flash on their smartphones.[4] Many homepages were not capable of displaying their content on iPhones and so they had to deploy their programs to JavaScript instead. To be able to program in a proper way, programmers started to build frameworks based on JavaScript. Most of these programmers were specialist in a different programming language and so they were used to different styles. Some of these new frameworks are no longer available and some have been growing over the years. It is hard to choose the right framework in the big pool of the internet. The top four JavaScript MVC frameworks are AngularJS, Ember.js, Backbone.js and KnockoutJS.

[2] Compare. Microsoft Developer Network
[3] Compare: Q-Success (2015)
[4] Compare: Jobs (2010)

As discussed in chapter MVC FRAMEWORK it is easier to program a software using a MVC framework. Therefore it is necessary to choose one of the existing frameworks. In the following a comparison of the top four JavaScript MVC frameworks will be conducted by defining a set of criteria and giving each criteria weighted points. In order to get a winning framework all points will be added together.

2.2.2 Set of criteria and rating definition

At first it is necessary to create a set of criteria to be able to rate which framework is the best. The choice on the criteria was made by different sources. [5][6][7]

1. What are the capabilities of the framework?
2. Does the framework have references? Are known projects using it?
3. How mature is it? How often occur bugfixes?
4. Is there a documentation available? Are there tutorials?
5. What is the size of the framework?
6. Is there a great community?
7. Are there any license restrictions?
8. Are there difficulties when installing?
9. Does it fit my style of programming?

The frameworks are being compared based on the criteria and given points. The points get a weight for an overall result. A rating with points for the capabilities is hard to define for each framework, because of dependencies for each project. A developer has to choose which features are necessary for the program he is going to build. Therefore there will be no rating for this criteria. The last criteria, the programming style, will neither be rated because it is a personal opinion and cannot be rated neutral. The least important criteria are is the maturity and therefore it only gets a factor of one. The following criteria have about the same importance and get a coefficient of two: references, size, license, installation. The most important criteria get a factor of three: documentation and community.

[5] Compare: Symfony (2015)
[6] Compare: tutsplus (2009)
[7] Compare: Safari Books Online (2013)

2.2.3 Set of criteria and rating execution

1. What are the capabilities of the framework?

AngularJS: The main feature of AngularJS is the data binding. It is very easy to manipulate the HTML code. Using JavaScript the programmer needs to change the DOM elements manually for each function. But AngularJS supports a two way data-binding that synchronizes the code between the DOM and the model. Another feature is the template system. AngularJS does not require the developer to change the templates by the JavaScript method `innerHTML`. It even supports a loop to create a table or similar for a set of data. Directives are a way to create custom HTML tags and add functions to those. So the developer is able to create his own tags for certain needs.[8]

Backbone.js: Backbone.js offers a RESTful-JSON-interface for a performant and maintainable connection between the client and server. All the data with stored in a key-value method.

Ember.js: One of the core features of Ember.js is the routing of objects by the URL. Ember.js uses Handlebar.js to handle templates for the user interface and they are automatically updated.

KnockoutJS: There are several features that are included in Knockout. One of them is the declarative bindings. Certain data can be linked to the UI using DOM.[9] Example:

```
Today's message is: <span data-bind="text: myMessage"></span>
```
[10]

Another feature are the automatic UI refreshes when data model's state changes. As the other frameworks use templates, so does Knockout too.

2. Does the framework have references? Are known projects using it?

LinkedIn.com, Soundcloud.com or Walmart.com belong to the portfolio of Backbone.js.[11] The other frameworks do not have any global top 50 websites as a refer-

[8] Compare: Ruebbelke (2012)
[9] Compare: Knockoutjs (2015a)
[10] Compare: Knockoutjs (2015a)

ences. But there is a changing trend. The figure 2 shows the progress of the usage by the global top 10,000 websites.

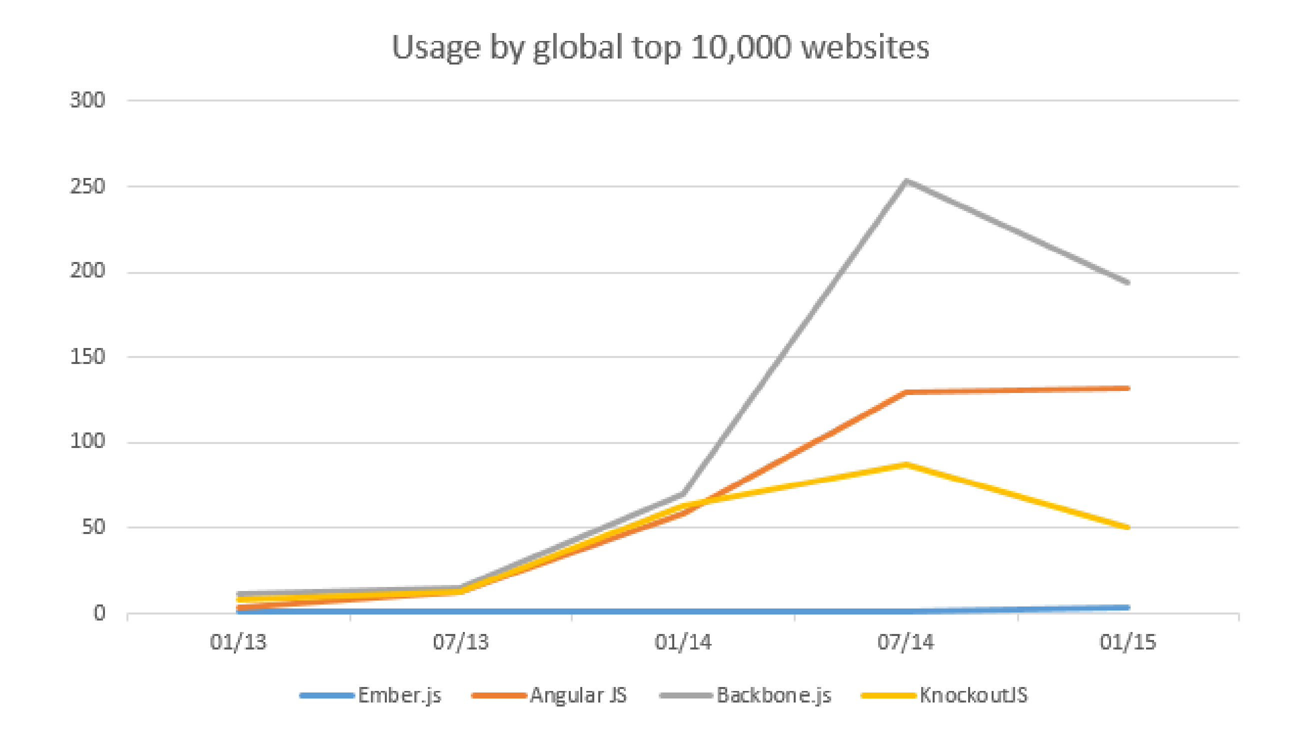


Figure 2: Usage by global top 10,000 websites

Ember.js has only three websites in the top 10,000 websites that are using it. KnockoutJS and Backbone.js are losing market shares from July 2014 to January 2015. AngularJS was able to keep and even raise their amount of websites a little bit. All in all the graph can show that Backbone.js is the leader right now but in the near future it may get passed by AngularJS. Backbone.js can compensate their losing market shares with the good portfolio.

Points:

Ember.js - Backbone.js • • • AngularJS • • KnockoutJS •

3. How mature is it? How often occur bugfixes?

The oldest of the four competitors is AngularJS, which development started in 2009. KnockoutJS and Backbone.js followed a year later in 2010. Ember.js is the youngest competitor with a release in late 2011. To be able to check on bugs it is useful to read the changelogs and see when the last bugs were fixed. Backbones.js last bug-

[11] Osmani, A. / O'Reilly, M. (2013), p. 2

fixes were made with version 1.1.2 on February 20th, 2014.[12] AngularJS newest release 1.4.0-beta.1 contains some bugfixes from January 20th, 2015.[13] Ember.js last bugs were fixed on December 8, 2014.[14] On August 12th, 2014 KnockoutJS released its last bugfixes.[15] All in all it is hard to tell if it is better that the last bugfix was made a long time ago or just a few days ago, because a longer period without a bugfix can either mean that there are no bugs anymore or nobody is checking on those anymore. So a recent bugfix does not need to be an indication for a bad software but it can indicate a big community that is taking care of the software. Only Backbone.js gets one point, because its last update or upgrades are almost a year ago.

Points:

Ember.js • •   Backbone.js •   AngularJS • •   KnockoutJS • •

4. Is there a documentation available? Are there tutorials?

All four competitors have a documentation on their website and all of them are offering references for the functions and features. Ember.js does not only offer a reference for the code but also has a built-in live preview in their tutorials. Except for tutorials by Ember.js itself there are only a few ones by third party websites. AngularJS offers videos and tutorials on the website and there are many third party websites that allow you to do a tutorial on learning AngularJS. KnockoutJS has an online tool for learning knockout which is really easy to understand for beginners. The code references are also available on their homepage. There are many third party homepages that also offer tutorials, but not as many as AngularJS or Backbone.js have. Backbone.js has a very good reference on site, but no tutorials. You need to find tutorials on third party pages, but there are a lot out there. All in all AngularJS documentation is the best and biggest one. Also because it has been continuously growing in the last years.

Points:

[12] Compare: backbonejs (2015)
[13] Compare: Github (2015a)
[14] Compare: Github (2015b)
[15] Compare: Knockoutjs (2015b)

Ember.js •   Backbone.js • •   AngularJS • • •   KnockoutJS • •

5. What is the size of the framework?

The size of the framework can be very important for some users. Even nowadays with mobile speeds above 100Mbit per second. An example can be that a client is online with his cellular phone in an area with a low internet speed. Therefore the whole page can be loaded quicker if the size of the framework is low.

Framework	Ember.js	Backbone.js	AngularJS	KnockoutJS
Size	348KB	7KB	45KB	22KB

Table 1: Size of the framework files

The table 1 shows the different sizes of the frameworks. The sizes are collected by the minimized files.

Points:

Ember.js -   Backbone.js • • •   AngularJS •   KnockoutJS • •

6. Is there a great community?

There are several ways to identify the size of a community. One way is to compare the amount of results an online search delivers:

Searchterm	Ember.js	Backbone.js	AngularJS	KnockoutJS
Results Google	538,000	769,000	16,800,000	544,000
Results Yahoo/Bing	1,970,000	4,010,000	2,610,000	4,780,000

Table 2: Search engine results based on the name of the framework as the searching word

AngularJS has the most results using Google. The reason for that may be that it is a Google product. KnockoutJS is on rank one using Yahoo or Bing search engine, but

some search results eventuate from the term knockout, so Backbone.js and KnockoutJS should be on rank one using Yahoo or Bing. The loser of this comparison is Ember.js. Having not many results on a search engine means a low action and discussion about this product. So it will be harder to find an existing community.

Another way to identify the size of the community is to check on the amount of registered users at meetup.com, an online portal where people with same interests can discuss on a certain topic.

Framework	Ember.js	Backbone.js	AngularJS	KnockoutJS
Members	22,695[16]	34,306[17]	76,263[18]	Not available

Table 3: Amount of members for each framework at meetup.com

This inquiry shows that AngularJS is the top of the 4 competitors, with twice as many members as Backbone.js which is on position two.

Both ways of defining the strength of the community together result in the following placing of the points.

Points:

Ember.js • Backbone.js • • AngularJS • • • KnockoutJS •

7. Are there any license restrictions?

All four frameworks are published under the MIT License.[19] The MIT License has its origin at the Massachusetts Institute of Technology and is a free software license. It allows a use of the software "without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions: The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software."[20]

[16] Compare: Meetup (2015a)
[17] Compare: Meetup (2015b)
[18] Compare: Meetup (2015c)
[19] Compare: Selle P. / Ruffles T. / Hiller C. / White J. (2014): p.13 / 37 / 89, KnockoutJS (2015)
[20] Compare: Open Source Initiative (2015)

As a result of no differences in the licenses there is no need to grant points to a framework.

8. Are there difficulties when installing?

Installing a JavaScript framework is in most cases done by simply downloading the files and integrated them by a `<script>` tag. But sometimes it is necessary to add some settings in the main layout file or include some other libraries that are dependencies to the framework. Backbone.js does not need any settings, but has dependencies to Underscore.js and jQuery which need to be downloaded seperatly.[21] AngularJS and KnockoutJS have neither dependencies nor necessary settings.[22] Ember.js also do not need any settings but it also has dependencies, jQuery and Handlebars.[23]

Points:

Ember.js • Backbone.js • AngularJS • • KnockoutJS • •

9. Does it fit my style of programming?

This is an important question for each programmer and there is no right or wrong. In order to be able to choose a framework, it is necessary to try out each framework and see which one is the best for the own purposes.

2.2.4 Conclusion

	Weighting	Ember.js	Backbone.js	AngularJS	KnockoutJS
1: capabilities	-	-	-	-	-
2: references	2	0 (0)	6 (3)	4 (2)	2 (1)
3: mature	1	2 (2)	1 (1)	2 (2)	2 (2)
4: documentation	3	3 (1)	6 (2)	9 (3)	6 (2)

[21] Compare: Selle P. / Ruffles T. / Hiller C. / White J. (2014): p.13
[22] Compare: Selle P. / Ruffles T. / Hiller C. / White J. (2014): p.37, KnockoutJS (2015)
[23] Compare: Selle P. / Ruffles T. / Hiller C. / White J. (2014): p.89

5: size	2	0 (0)	6 (3)	2 (1)	4 (2)
6: community	3	3 (1)	6 (2)	9 (3)	3 (1)
7: license	2	-	-	-	-
8: installation	2	2 (1)	2 (1)	4 (2)	4 (2)
9: style	-	-	-	-	-
Total		**10**	**27**	**30**	**21**

Table 4: Set of criteria scorecard

The result of the comparison can be seen in table xxx. AngularJS won the comparison with 30 points against BackboneJS (27), KnockoutJS (21) and EmberJS (10). Overall AngularJS seems to be the best choice at the moment and the near future. Ember.js and KnockoutJS are only alternatives to the two big frameworks. BackboneJS seems to be losing against AngularJS, because of the global impact of Google and their image.

3 Technologies used

This chapter provides basic information of different languages/technologies that you need when developing a Web application.

3.1 HTML

Hypertext Markup Language is probably the most important and first learned programming language when it comes to the development of Web applications. HTML defines the architecture/structure of the website (the view you can see). HTML-files consist of normal text the visitor of the website will see. The important part is the additional code which defines the structure and determines what part of the text is a heading or a paragraph or something else. Therefore the content of an HTML-file is ordered into different HTML elements. These elements are marked with specific tags that define the type of the element. Almost all of them are marked by an initiatory and a closing tag thus enclosing the part of text they are referring to.

```
<!DOCTYPE html>
<html>
  <head>
    <meta http-equiv="content-type" content="text/html; charset=utf-8">
    <title>Elemente und Tags</title>
  </head>

  <body>
    <h1>HTML - die Sprache des Web</h1>
  </body>
</html>
```

Figure 3: Sample HTML code[24]

Figure 2 shows an example of how this structure is implemented. Within the body of the file we have a heading which begins with the **<h1>** tag and ends with **</h1>**. Yet there are also elements which have no content thus consisting only of one single tag. E.g. a line break is created by the single tag **
** and needs no closing tag.

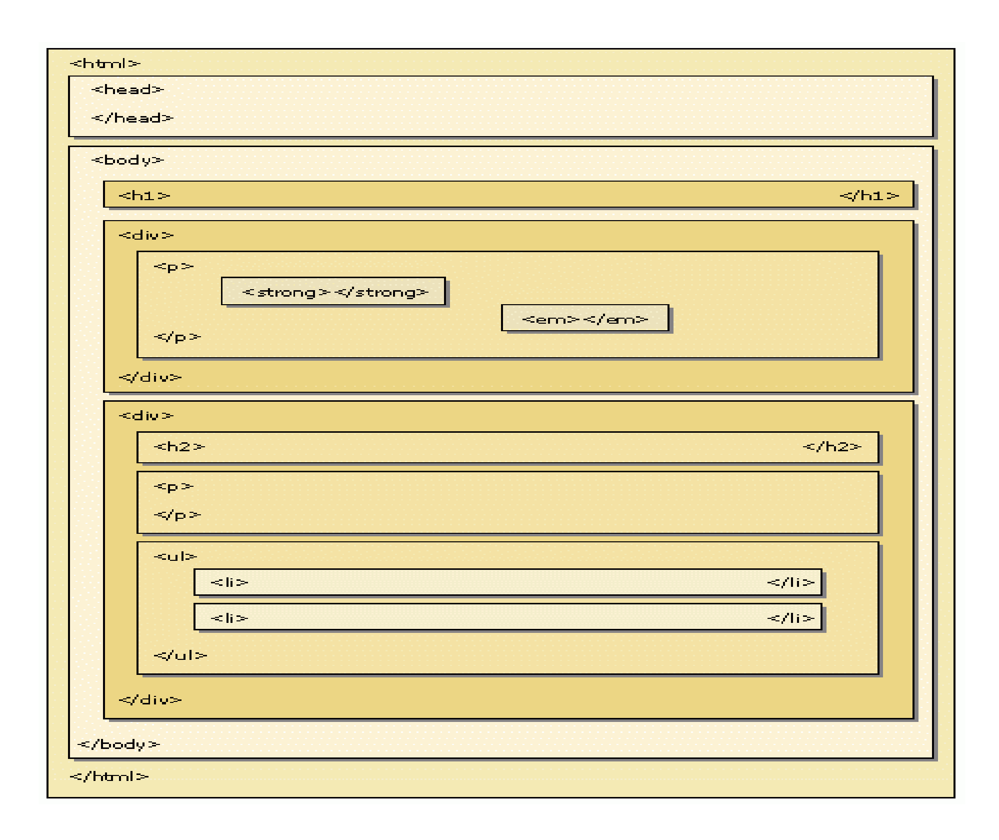


Figure 4: HTML structure[25]

Figure 3 displays the logical structure of an HTML-file. It nicely illustrates how the several elements are built together. They are nested within each other and are forming a hierarchical structure, also called structured markup.

[24] From: selfhtml.wiki (2015a)

[25] From: Lee Underwood (2015)

Initiatory tags as well as standalone tags can contain additional information/attributes. This is especially important when coming to the practical implementation with the usage of the AngularJS framework.

There are different types of attributes for HTML elements:

- Attributes that only predefined values can be assigned to, e.g. **<input type=”text”>**. Possible values for the attribute “type” are limited to a given list of existing ones.
- Attributes with open value assignment, but need to fulfil a certain data type or convention. An example for this would be to extend the previous input element with such an attribute: **<input type=”text” maxlength”10”>**
- Attributes with completely open assignment without any conventions, e.g. **<p title=”anything you like”>**
- Single attributes without value assignment: **<input type=”text” readonly>**

With the new HTML5 the development of HTML code and websites has become even more structured and easier. Now there are different elements like <header> and <footer> and you no longer need to use <div> for it and assign different IDs or classes. Besides that there are more elements like <aside> and the new <nav> that make HTML5 a really easy to use and attractive programming language.[26]

3.2 CSS

In the previous chapter it was all about HTML and defining the structure and architecture of the website and the various elements. When you want to display the HTML in your browser with just the structure the elements will just show one below the other as they are ordered in the HTML file. The display will be according to the standard settings of the browser, probably like black and white and headings will be bold. But this doesn’t fulfil the visual requirements of a properly designed website.

[26] Compare: selfhmtl.wiki (2015b)

The solution is Cascading Style Sheets (CSS). CSS is a language primarily developed to complement HTML. It allows any formatting of single HTML elements. So for example you can define, that class 1 headings always have a certain colour, size, font and so on. Similarly you can set the attributes for every HTML element and how it shall be displayed. HTML is responsible for the structure of the elements and CSS complements it with the visual design.

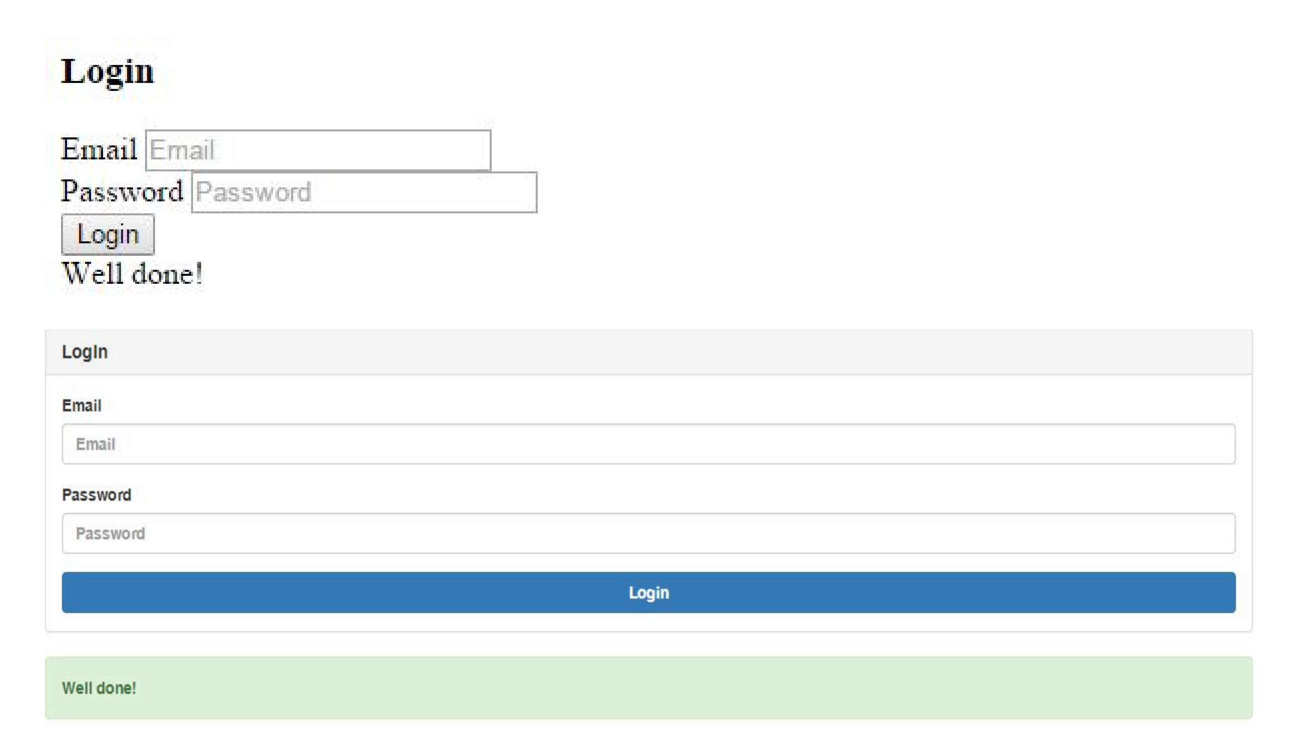


Figure 5: Impact of CSS

In Figure 4 you can see the difference between an HTML without the use of CSS (top) and with the use of CSS (bottom). Without the visual formatting the website simply just would not look appealing and users will not enjoy using them.

Therefore CSS is both a simple and clean way to visually upgrade your HTML file. CSS enables a central formatting in a separate CSS-file that refers to the elements in the HTML and defines their attributes. Any number of HTML files can refer to a central Stylesheet and provide a consistent design e.g. for bigger projects and corporate design.

In a Cascading Stylesheet attributes are ordered in a set of rules. A rule consists of a selector or a group of them, followed by an area in which values are assigned to the certain attributes:

Selector { Attribute1: Value; Attribute2: Value; }

A Stylesheet can have only one or many hundreds of rules.

As already mentioned the most common and clean way to implement CSS is to have a separate file where your rules are defined. In order to use this Stylesheet for your HTML you therefore have to embed it into your HTML file and build in an link:

<link rel="stylesheet" href="mystylesheet.css" type="text/css">

The attribute **rel** describes the relation type and that a stylesheet is embedded. The **href** attribute is a concrete reference to the certain CSS-file.

In contrast to the presented method you can also implement CSS directly in your HTML file. Either you can define formatting in a central style element within your HTML code or you directly modify single elements with the style attribute. But handling your style within the HTML file is not that common since many advantages are lost using so called "Inline-Style". The formatting is bound and limited to this one file and cannot be changed at a central place thus lowering flexibility during development.[27]

3.3 JavaScript

JavaScript is not a certain part of HTML but is an own programming language that can be used by programmers to optimise their websites. Similar to CSS JavaScript can be implemented directly in the HTML file or in a separate file. It is a powerful tool which can be used to write some easy routines up to building complex frameworks.

JavaScript is not easy to learn but easy to use because it is running in a so called "sandbox" which limits the languages options and cuts off certain functionalities of other, bigger programming languages. Especially the ability to read and write into files is restricted in order to prevent attacks to users who are using websites supported by JavaScript.

[27] Compare: selfhtml.wiki (2015c)

Whilst HTML provides the structure and CSS the presentation JavaScript handles the interaction with the user. Similarly as CSS JavaScript aims to improve the usability of an application

Since JavaScript is running in the browser of the client when a website is loaded or certain events take place an area of functions that are server and data related can't be provided by JavaScript.

As mentioned above JavaScript can be built directly into the HTML file or be stored in an own file. When implementing the code into your HTML file you write the code in the <script> element:

```
<script>
    alert("Hello World!");
</script>
```

Both for CSS and JavaScript however the better method is to store the code in a separate file and just add a reference to the HTML file. The script is then running as if the code was written directly into the HTML:

```
<script src="MyScript.js"></script>
```
[28]

3.4 AngularJS

AngularJS, developed by Google, is the framework technology used in the practical part of this work. It uses JavaScript in combination with HTML to build client side web applications. Its focus lies less on the JavaScript code but rather on the describing abilities of HTML. The framework was developed with the idea of how HTML would look like if it was designed for application development.

AngularJS enables the use of HTML and lets you extend the syntax to define own HTML elements and attributes in order to express your applications components clearly. The framework's data binding and dependency injection get rid of much of

[28] Compare: selfhtml.wiki (2015d)

the code that would be needed elsewise. It can be seen as a complete client side solution.[29]

4 The Project

4.1 Background

The main purpose of this paper is to analyze and compare several JavaScript frameworks that rely on the MVC design pattern. As it turned out in the last chapter, AngularJS perfectly fits the needs for a Web-Application that is to be developed in the context of this project. However, when the project started, the functionality of the Web-Application was still unclear. This changed with a little research of trending IT technologies.

Ebooks experienced a great increase in number of users over the last few years. Not only are they used in designated devices, a lot of users simply view them on the web. With the use of responsive web designs, the portability is still ensured. But there is still one thing that you can do with traditional books, which you can't with ebooks. You can't interact with an ebook in a way that you can with a traditional book.

Regarding novels or short stories this issue might seem insignificant. But educational books often include the possibility to finish homework assignments directly in it. The goal of this project is to remove this issue by implementing an interactive ebook reader that includes the functionality to answer questions or finish homework assignments within the ebook layout. Including a user management system, this application might be used by schools to decrease the effort in managing and marking homework assignments as well as reduce the amount of paper that is used by schools.

In the following sections the overall architecture as well as technical features will be explained. Additionally, the code snippets for the main functionality that relies on AngularJS will be shown.

[29] Compare: AngularJS (2015)

4.2 Architecture

This section explains technical features of the Web-Application as well as the overall architecture. Since the background of this project is to use MVC-Frameworks, we will first look into the separation of data, business-logic and presentation of the application.

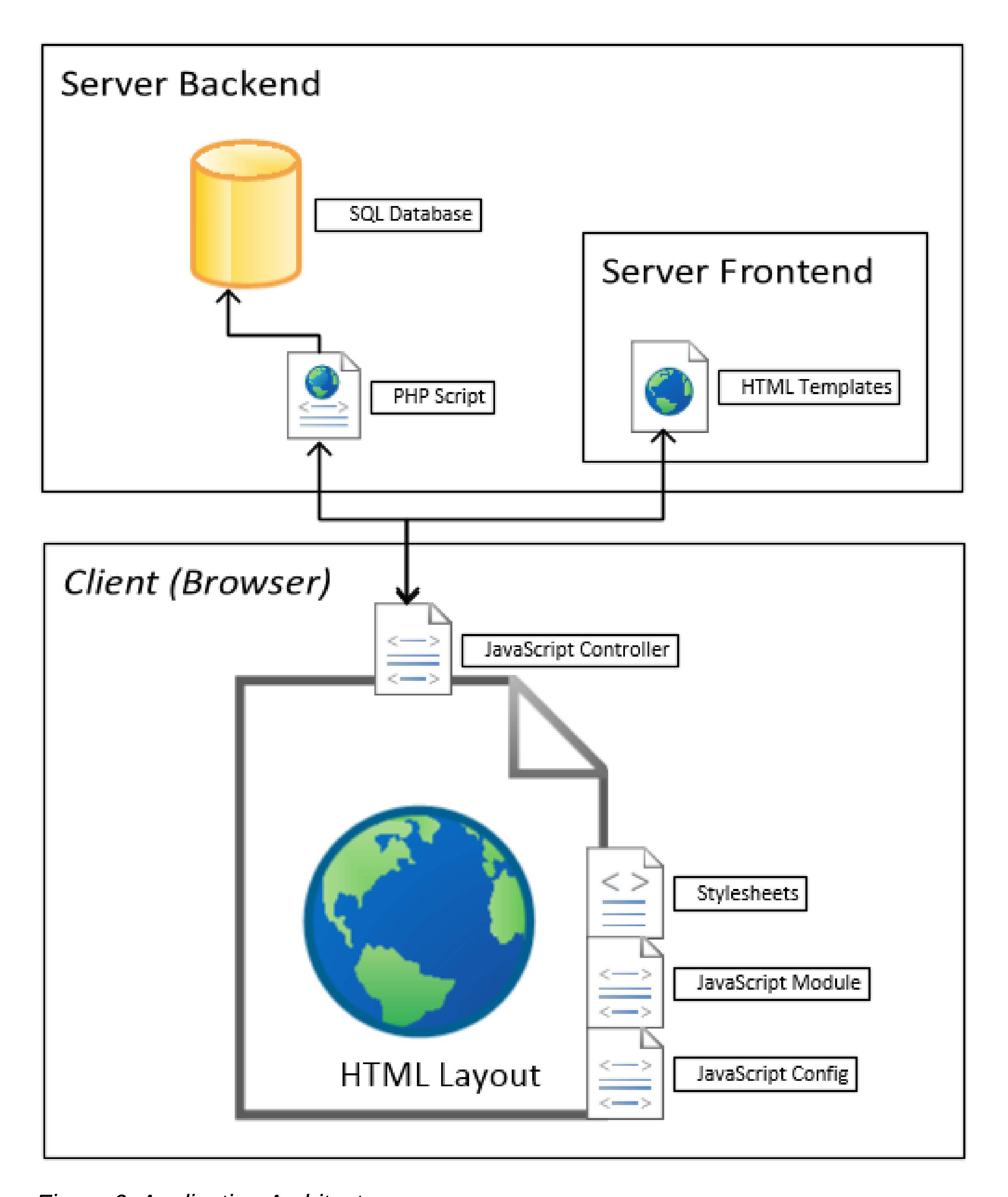


Figure 6: Application Architecture

Figure 6 shows the basic application architecture at runtime. The process of creating a View for the user is relatively easy. Once a user opens the website, the HTML layout and all referenced JavaScript-files as well as style sheets are loaded into the browser. After that, the application is started by constructing the first AngularJS module in *JavaScript Module*.

After configuring (*JavaScript configuration*) and constructing the remaining controllers (*JavaScript Controller*) the actual view is created by a controller loading model data from the database, binding the data to an HTML template loaded from the Webserver and loading the result into the HTML layout. If the user navigates through the website, the only things that need to be loaded into the browser are the different HTML templates as well as the application data. The HTML basis stays the same throughout the application lifecycle.

The application lifecycle starts with the user opening the website and ends with him/her closing the browser tab. To explain the different features and underlying processes, the following section will guide through such an imaginary application lifecycle.

4.3 Application lifecycle

4.3.1 Loading the website

As mentioned, after a user opens the website the HTML foundation is loaded into the browser. It contains only a navigation bar and a container for the HTML templates that will be requested by the controllers. Furthermore it loads all the referenced JavaScript controllers containing the business logic and the referenced style sheets.

```
<html ng-app=”app” ng-controller=”AppCtrl”>
<div id=”templateContent” class=”content” ng-view>
```

The *ng-app*-directive in the opening HTML-tag connects the HTML with the corresponding AngularJS-application *app*. By doing this, the AngularJS framework identifies the module that needs to be started and ensures that it executes the functions *app.config()* and *app.run()*. The *ng-view* directive on the other hand binds the separate HTML templates to this container. So every time an HTML template is loaded it will be inserted into the <div> with the ID *templateContent*.

```
app.config(['$routeProvider', function($routeProvider, $win-
dow, $rootScope, $q) {
     $routeProvider.
          when('/home', {templateUrl:'templates/home.html'}).
            //rest of the routes and authorization management
                  otherwise({redirectTo:'/home'})
```

The function *app.config()* implements the routing functionality by accessing the AngularJS provider *$routeProvider*. This way, for every resource the user is authorized to access corresponding HTML templates and controllers are specified. The authorization management will be explained in one of the following sections.

```
app.run(function($rootScope, $location){
        console.log('app is running');
})
```

After processing the code in *app.config()*, *app.run()* is executed as the last activity before the controller takes over. In this example it simply logs a status message ("app is running") to the browser console.

Up to this point the view is not yet created, leaving the user with a navigation bar at the top of the page. However, this will change with the controller being executed and the template being loaded into the container after an authorization check is performed.

4.3.2 Authentication and Log-In

The authentication process is implemented using JSON Web Tokens (JWT) that are stored in the local browser storage. Since the user is not yet logged in there is no token available. A simple check for an existing token is performed in the *app.config()* function.

```
when('/home', {
                templateUrl:'templates/home.html',
                resolve : {
                      //Code for initializing check
```

```
if(angular.isDefined($window.sessionStorage.token))
                     { //Code for user with existing token }
else
                     { //Redirecting users with no token to
login }
                }
//Code for the remaining routes
})
```

Angular.isDefined($window.sessionStorage.token) scans the local browser storage for such a token, returning true if it can or false if it can't find one. Users with no token are redirected to the login page, hence the *login*-controller is activated, loads the HTML template and completes the first view.

At this point, the user has to insert the login data and send it by clicking the login button. This triggers an event that is handled inside the *login*-controller. When the event gets triggered, the function *submit()* manages the communication between client and server.

```
$http.post(url, $scope.user)
        .success(function (data,status, headers,config){
        //Code for setting the token and redirecting the user
             })
        .error(function (data,status,headers,config){
        //Code in case the login fails
             })
```

The $http service sends an HTTP POST request to the server that contains the user data. If the data is valid, the server answers with an HTTP response containing the status code "200 OK", meaning the user is authorized to access several resources and will be redirected to the *home*-URL. The response also contains the token that is stored in the local browser storage. In case of an invalid input by the user, the server will respond with a status code of "401 Unauthorized" and the user has to try again. Since the mechanism of registering a user uses the same principles, it will not be mentioned in this paper.

Following this mechanism, every HTTP request in the controllers integrate the token into their HTTP headers. This way the server can authorize the user on every incoming request by validating the token with the secret server key.

4.3.3 User Account

So far, the user opened the website and logged into his account. That process redirects him to his user profile, hence activating the *home*-controller which loads the *home*-template onto the screen. Now, the first step of the controller is to load the user specific information from the database using an HTTP request.

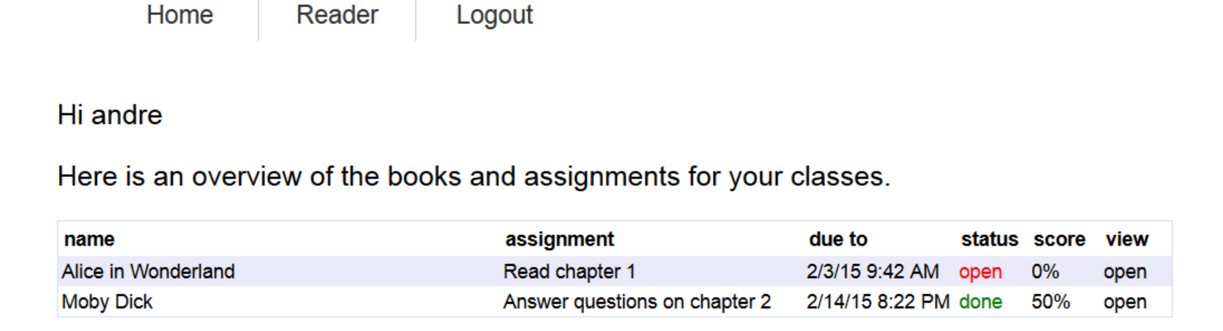


Figure 7: User Account

As one can see on Figure 7 , all the information is displayed in a structured manner, giving the user a good general view over his account. These information include the homework that needs to be done and the due dates, as well as the score of his finished assignments. On the right hand side he can find links that lead directly to the actual ebook reader.

4.3.4 Interactive Ebook Reader

Let’s assume, the user wants to review his questions in Moby Dick. He clicks on the link and the *reader*-controller takes over, hence loading the reader-template. This

template contains an <iframe> that imports the ebook reader with the important JavaScript.

Since it would be great time investment to developean own ebook reader, this application includes the open source project *futurepress.* This ebook reader supports the mostly used .epub format. These open standard containers comprise of xml or xhtml files for the content, as well as meta information in xml format.

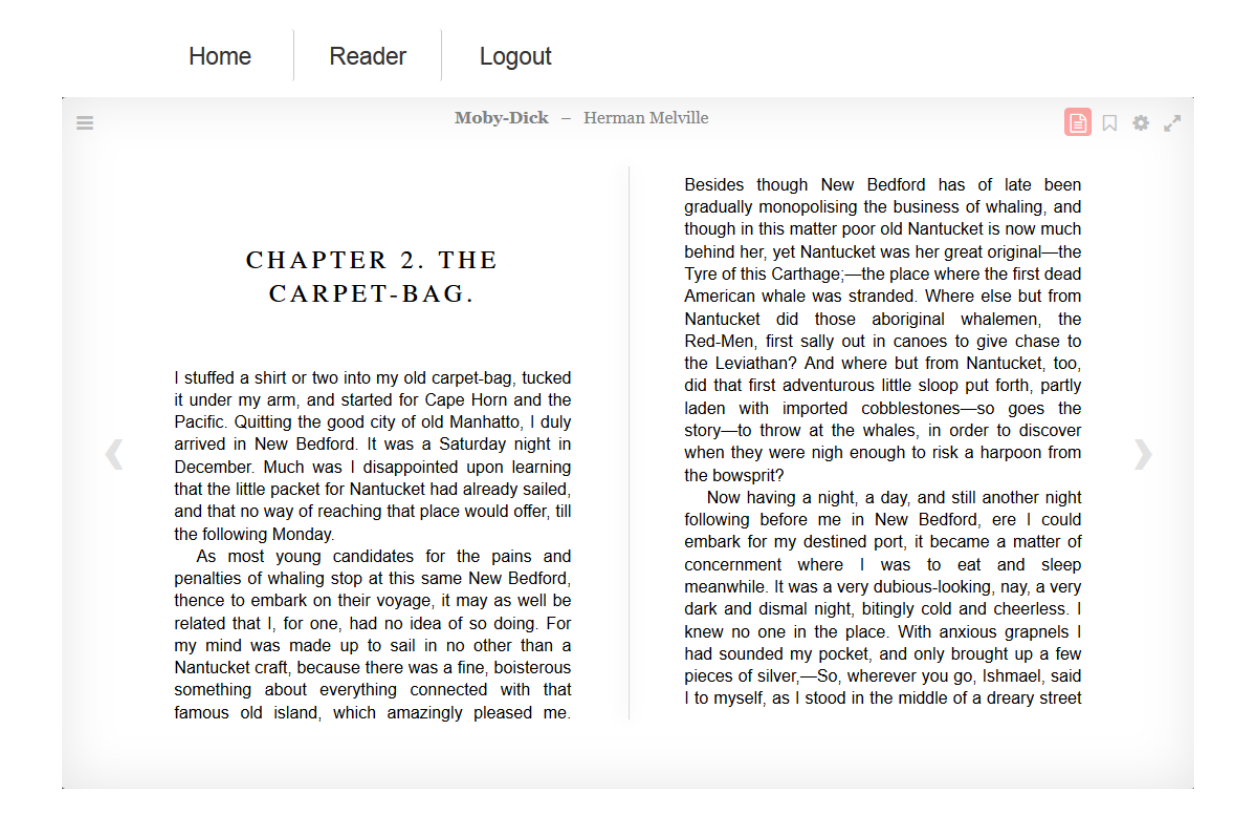


Figure 8: Ebook reader

Figure 8 shows the ebook reader in action. On the top one can still see the navigation bar and, on the top right hand corner, an additional menu for the reader. The red icon indicates that there is home to be done at the end of this chapter, in this case chapter two. If there was no homework, the icon would appear white.

By clicking on the red icon, a pop-up window appears with the corresponding questions. Now the user has the opportunity to answer the questions and later review the score on his user account.

4.4 Additional Assumptions

4.4.1 Data Model

The application of this project relies on a very simple data model with only four tables.

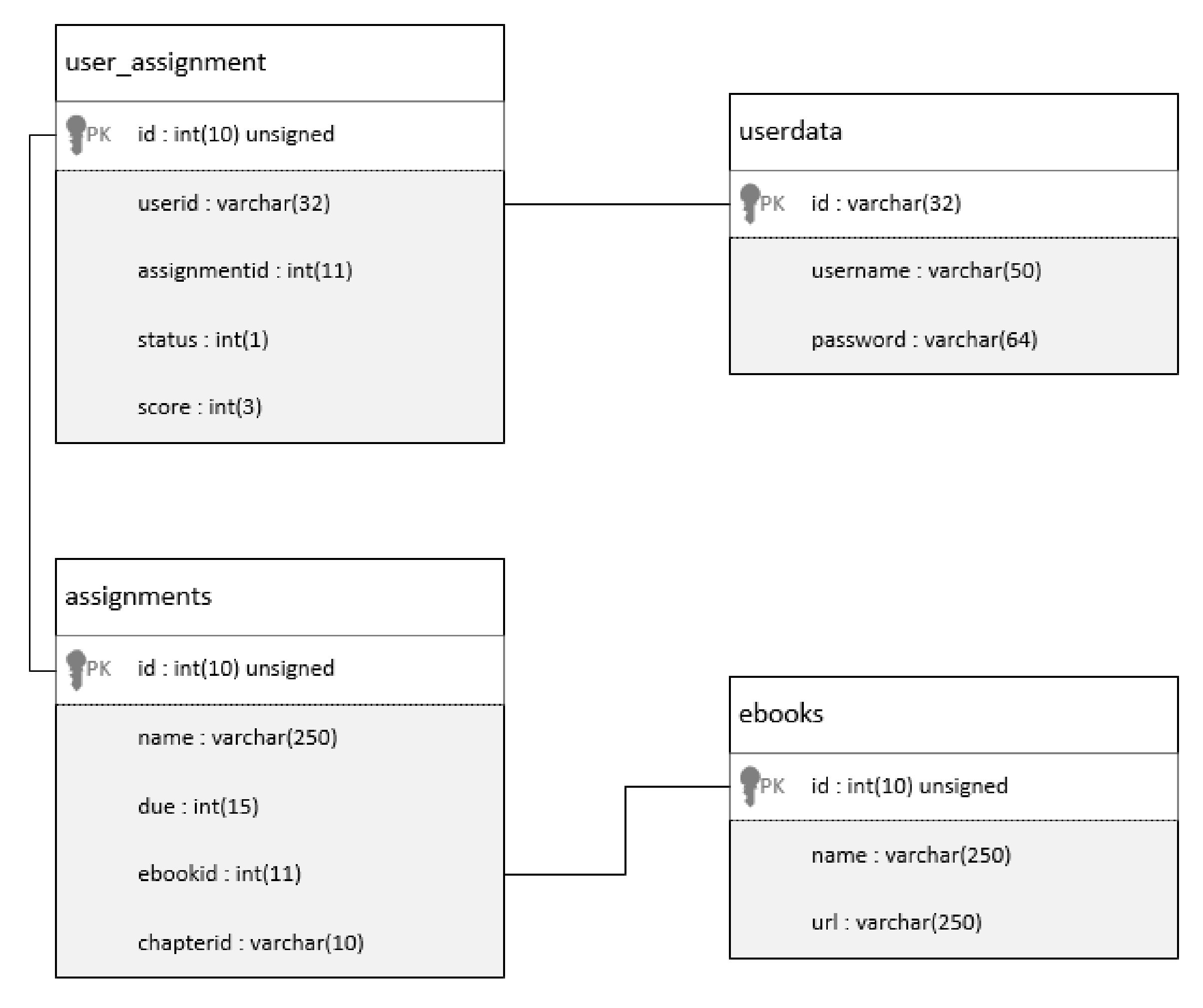


Figure 9: Data Model

Figure 9 shows the data model that has been used in this project. *Userdata* contains all the user information that is needed to log in, *user_assignment* connects the user to his assignments. Finally the user gets the ebook data that is needed via *assignments* from *ebooks*.

4.4.2 Project Scale

The original task in this project was to develop an interactive ebook reader that implements a way to answer questions or finish assignments online. Since there would be no sense in developing that with an MVC based framework, a user management system was developed around that. The resulting system can be used by schools to manage homework assignments ordered by class membership. In order to do so, the current system needs to implement a user management that is based on different authorization levels. Authorized users like teachers can then upload ebooks where users with lower-level authorization have access to. Also users have to be categorized into class membership to prevent unauthorized access to test results.

Furthermore, to deal with increased usage of mobile devices, the user interface needs to be restructured to a responsive design. By doing that, students get the opportunity to finish homework assignments anywhere at any time. Teachers on the other hand get a structured homework management system that can rapidly reduce their time invested into scoring the homework results.

Lastly, an ebook configuration system may be implemented into the existing project. By inserting the necessary JavaScript functionality into an Epub file, every school book could be made compatible to the interactive ebook reader. A school could save a lot of expenditures by switching to such a paperless homework management system and, at the same time, save paper to help the environment.

4.4.3 Security

Security mechanisms in this project rely on the JSON Web Token method. A single token that is stored on the client side gets forwarded to the server in the HTTP headers of every single HTTP request. On the server side, this token has to be verified by simply decoding it with the secret server key.

At this point, no encryption mechanisms for the HTTP requests has been implemented. This means, that the user data as well as the token are not protected from illegal access while the package is sent through the network. By using HTTPS instead of HTTP, this problem could be solved. Additional safety could be accomplished by binding the user's token to his IP address with every log-in. Such a double layer se-

curity system would provide enough safety for the users and their corresponding data.

Without these safety measures, tokens would not be safe from being hijacked over a Local Area Network. By obtaining a user's token and inserting that token to HTTP requests, an attacker would have access to all the user's data. Even if the user management system does not contain any credit card information or other critical data sets, providing the user with a safe way to access his data is critical in any application.

4.4.4 MVC Design Pattern

Prior to implementing an example application, the first task in this project was to examine several JavaScript frameworks that provide a way to develop an application following the MVC design pattern. In this case the AngularJS framework was chosen due to a couple of reasons explained in the previous chapters. But how did AngularJS and the MVC design pattern in general benefit the development process of the application?

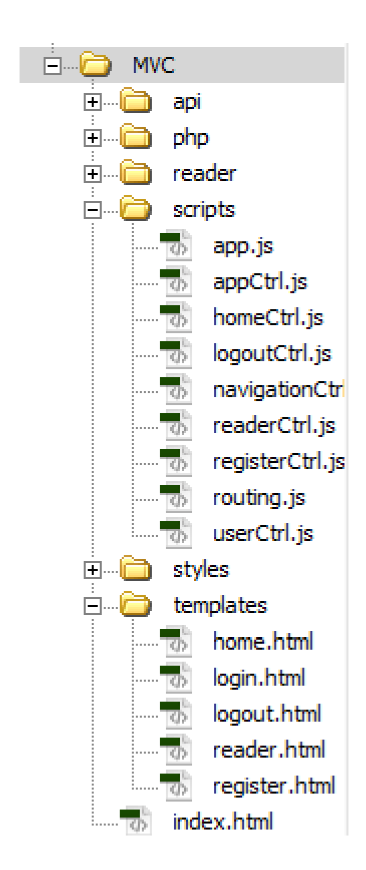


Figure 10: Directory Structure

Figure 10 shows the directory structure of this project. Following the MVC design pattern simplifies the process of making changes to any resource. E.g. if a routing procedure has to be changed or added, a few lines of code in *routing.js* would be enough. Another profit of using a structure like this is that some of the basic code modules can be reused in future projects. This would apply for routing procedures as well as log-in and registering mechanisms.

So far, the scope of this project was limited to implementing the interactive ebook reader and a user management system. Even in such a project the separation of design, business logic and data had a great impact on the developing process. To develop and test a single code module was accomplished without having to implement it into the system. The same situation occurred when shaping a new HTML template.

The AngularJS framework offered all the functionality needed for the development process. Even people with only a little knowledge background in JavaScript were able to develop whole code segments in a timely and structured manner and, since the single code modules were relatively small, the process of identifying and correcting mistakes was very easy.

In the last section an imaginary scale-up was performed on the project. In reality, the implementation of additional functionality as well as new designs could be achieved in a matter of days. Since the project is well structured and naming conventions were met, new developers would have it easy to integrate themselves into the developing process.

5 Conclusion

As was shown in the previous chapters, JavaScript and its frameworks gain more and more attention for a reason. Developing client side applications with a high rate of portability, functionality and security is as easy as never before. If these applications are based on the MVC design pattern, they can easily be tested and scaled-up. Furthermore, following these patterns increases the maintenance and reusability of existing code.

In this project, an example application had to be developed to test these features of client side JavaScript frameworks. The goal was to implement an ebook reader with the capability to obtain user interactions. Using an open source ebook reader that supports the famous .epub format, the application now includes a user management system as well as the functionality to manage homework assignments for students and teachers.

In my opinion, working with the AngularJS framework was easy to learn. The basic implementation of controllers and templates may be achieved within hours, even with only a little knowledge in programming and web design.

References

Literature

Osmani, A. / O'Reilly, M. (2013): Developing Backbone.js Applications

Selle, P. / Ruffles, T. / Hiller, C. / White, J. (2014): Choosing a JavaScript Framework, Bleeding Edge Press

Internet and Intranet

AngularJS (2015: What is Angular?, https://docs.angularjs.org/guide/introduction

Backbonejs (2015): Changelog, http://backbonejs.org/#changelog

Github (2015a): 1.4.0-beta.1 trepidatious-salamander (2015-01-20), https://github.com/angular/angular.js/blob/master/CHANGELOG.md

Github (2015b); Ember Changelog, https://github.com/emberjs/ember.js/blob/master/CHANGELOG.md

Jobs (2010): "Thoughts on Flash", https://www.apple.com/hotnews/thoughts-on-flash/

Knockoutjs (2015a): Knockoutjs Documentation, http://knockoutjs.com/documentation/binding-syntax.html

Knockoutjs (2015b): Knockoutjs Downloads, http://knockoutjs.com/downloads/

Lee Underwood (2015): The HTML Hierarchy: Thinking Inside the Box, http://www.htmlgoodies.com/beyond/article.php/3681551/The-HTML-Hierarchy-Thinking-Inside-the-Box.htm

Meetup (2015a): Meetup-Gruppen zu Ember JS, http://ember-js.meetup.com/

Meetup (2015b): Meetup-Gruppen zu Backbone.js, http://backbone-js.meetup.com/

Meetup (2015c): Meetup-Gruppen zu Angular JS, http://angularjs.meetup.com/

Microsoft Developer Network (2015): Model-View-Controller, https://msdn.microsoft.com/en-us/library/ff649643.aspx

Open Source Initiative (2015): The MIT License (MIT), http://opensource.org/licenses/mit-license.php

Q-Success (2015): Historical trends in the usage of client-side programming languages for websites, http://w3techs.com/technologies/history_overview/client_side_language/all

Ruebbelke (2012): 5 Awesome AngularJS Features, http://code.tutsplus.com/tutorials/5-awesome-angularjs-features--net-25651

Safari Books Online (2013): 13 Criteria for Evaluating Web Frameworks, https://blog.safaribooksonline.com/2013/10/14/13-criteria-for-evaluating-web-frameworks/

Selfhtml.wiki (2015a): HTML/Allgemeine Regeln/Textauszeichnung, http://wiki.selfhtml.org/wiki/HTML/Allgemeine_Regeln/Textauszeichnung

Selfhtml.wiki (2015b): HTML, http://wiki.selfhtml.org/wiki/HTML

Selfhtml.wiki (2015c): CSS, http://wiki.selfhtml.org/wiki/CSS

Selfhtml.wiki (2015d): JavaScript, http://wiki.selfhtml.org/wiki/JavaScript

Symfony (2015): 10 criteria for choosing the correct framework, http://symfony.com/ten-criteria

tutsplus (2009): 15 Most Important Considerations when Choosing a Web Development Framework



Untersuchung von Open Source Thin Client Produkten in Verbindung mit einer Citrix VDI-Umgebung

Schriftliche Ausarbeitung
im Rahmen der Lehrveranstaltung „Integrationsseminar“
für das Kompetenzzentrum Open Source (KOS)

Vorgelegt von

Maik Blümel, Sebastian Lutz,
Philipp Schreyer, Miriam Senger,
Helen Wrona

am 26.01.2015

Fakultät Wirtschaft
Studiengang WI – International Management for Business and Information Technology
WWI2012I

Inhaltsverzeichnis

Abkürzungsverzeichnis

LAN	Local Area Network
OSI	Open Source Initiative
OSS	Open Source Software
PC	Personal Computer
UMS	Universal Management Suite
VDI	Virtual Desktop Infrastructure
VM	Virtuellen Maschinen
VPN	Virtual Private Network
WAN	Wide Area Network

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

In Zusammenarbeit mit Wirtschaftsinformatik Studenten des 5. Semesters der Dualen Hochschule Baden-Württemberg in Stuttgart startet die .Versicherung ein Projekt zur „Untersuchung von Opensource ThinClient Produkten in Verbindung mit einer Citrix VDI-Umgebung“.

Bisher sind bei der .Versicherung eine Vielzahl von FatClients in Form von Personal Computers (PC) im Einsatz. Dies erschwert die Verwaltung der PCs und deren Systeme, da die Konfiguration und sonstiger Support nur vor Ort vorgenommen werden kann. Um dieser Herausforderung entgegenzutreten, kommen verschiedene Ansätze in Betracht. In diesem Projekt werden ThinClients in Virtual Desktop Infrastructure (VDI) Umgebungen untersucht. Diese sollen eine zentrale Verwaltung anhand einer Verwaltungskonsole ermöglichen. Dabei ist zu unterscheiden zwischen Open Source und kommerziellen Produkten. In dieser Arbeit liegt der Fokus jedoch auf Ersterem.

Das Ziel ist, eine Marktstudie zu den momentan am Markt befindlichen Open Source ThinClient-Produkten zu erstellen. Verschiedene Open Source ThinClient Betriebssysteme sollen hinsichtlich ihrer Einsatzbarkeit im VDI-Umfeld der .Versicherung evaluiert werden. Die VDI soll unter Verwendung von bereits bestehenden Virtualisierungstechnologieherstellern realisiert werden.

Zunächst werden die Grundlagen von VDI, ThinClient, Open Source und Virtual Private Network erforscht, sowie Beispiele anhand Best practices gegeben.
Ein Kriterienkatalog zur Evaluierung des besten Produkts wird erstellt und beschrieben. Dieser umfasst den Funktionsumfang (inklusive Verwaltungskonsole, Citrix- und VPN-Fähigkeit), die Verbreitung, Reife, das Lizenzmodell, die Einfachheit der Implementierung, sowie der dazugehörigen Gebühren. Diese Kriterien werden anhand von Kategorien priorisiert und erhalten jeweils eine Bewertung. Abschließend erfolgt die Evaluierung anhand einer Berechnung, in der die Ergebnisse der Bewertung und Priorisierung miteinfließen.

Abschließend sollen die Funktionen der ThinClient-Lösung anhand einer Testumgebung evaluiert werden. Diese soll ein vereinfachtes Szenario darstellen, innerhalb dessen sich der Funktionsumfang, vor allem aber die Administrierbarkeit erproben lässt.

2 Theoretische Betrachtungen

Im Folgenden werden die Grundlagen und die methodische Vorgehensweise betrachtet. Dabei dient der Grundlagenabschnitt zur Begriffs- und Technologieerläuterung, sowie zur Dar-

stellung von Best Practices. Im Abschnitt „Methodisches Vorgehen“ wird auf das Prinzip des Kriterienkatalogs eingegangen und das Bewertungsvorgehen erläutert.

2.1 Grundlagen

2.1.1 VDI-Technologie

2.1.1.1 VDI-Umgebung

Eine Virtual Desktop Infrastructure (VDI) zeichnet sich durch die Desktop-Virtualisierung aus, bei der die Daten und Programme von dem physischen, lokalen Endgerät des Anwenders abstrahiert werden. Der Desktop wird auf einem zentralen Server gespeichert anstatt auf der lokalen Festplatte eines Laptops oder Personal Computers (PC), sodass man ihn als virtuellen Desktop bezeichnet. Alle Programme, Anwendungen, Applikationen und Daten werden ebenfalls zentral im Rechenzentrum gespeichert. Der Anwender ist beim Arbeiten an seinem Desktop somit nicht mehr an ein bestimmtes Endgerät gebunden, sondern kann auf den virtuellen Desktop von jedem beliebigen Gerät zugreifen, das sich mit dem zentralen Server verbinden kann. Das sind beispielsweise Laptops, PCs, Smartphones sowie ThinClients. Der Zugriff kann über jedes beliebige Netzwerk erfolgen, beispielsweise einem Local Area Network (LAN), Wide Area Network (WAN) oder dem Internet, sodass der Anwender jederzeit von zu Hause und anderen Orten auf seinem Desktop arbeiten kann.[1]

Durch die Virtualisierung ist es möglich, verschiedene Betriebssysteme auf den virtuellen Desktops bereitzustellen, selbst wenn diese auf demselben Server laufen. Dies wird ermöglicht durch die Virtuellen Maschinen (VM), die einzeln partitioniert werden und den virtuellen Desktop abbilden.[2] Dies erlaubt dem Unternehmen, mit einem Server die unterschiedlichen Bedürfnisse der Anwender zu erfüllen. Der benötigte Desktop wird auf dem spezifischen Client des Anwenders bereitgestellt, was einen hohen Grad an Flexibilität ausmacht.[3]

[1] Vgl. Redondo Gil, C. u. a. (2014), S.37 f.
[2] Vgl. Niemer, M. (2010), S. 58
[3] Vgl. Liu, X./Sheng, W./Wang, J. (2011), S. 413

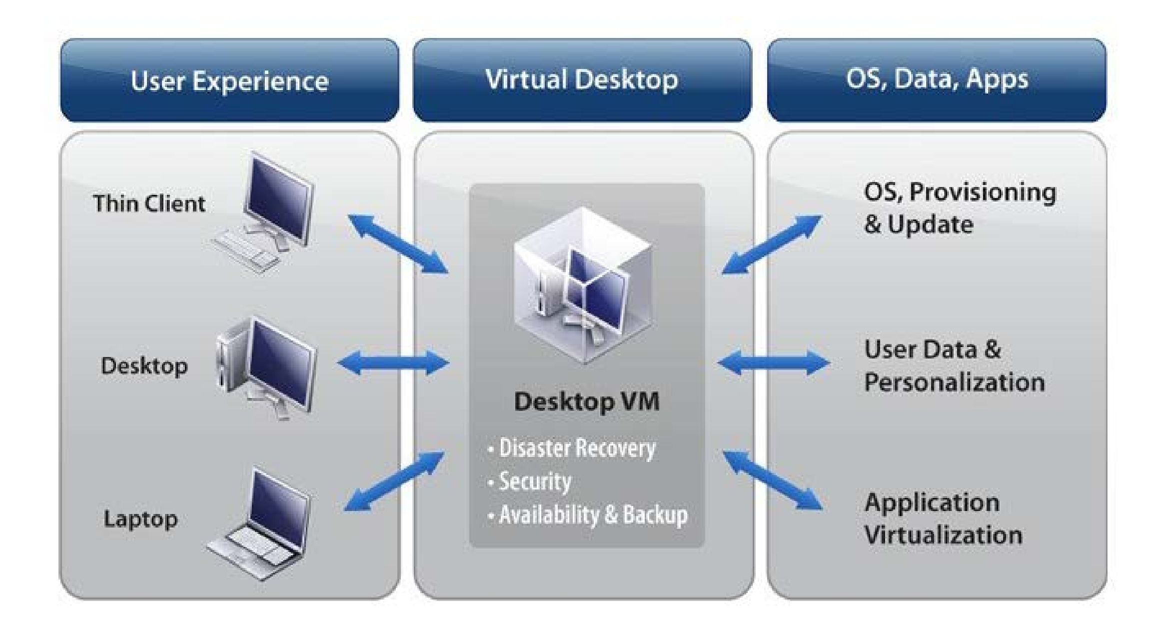


Abb. 1: Schaubild einer VDI-Umgebung[4]

Abbildung 1 veranschaulicht das Modell der VDI. Es zeigt, dass der Nutzer von seinem Endgerät auf den virtuellen Desktop zugreift und dieser wiederum auf die Daten und Applikationen zugreift. Zudem wird verdeutlicht, dass Updates und Backups auf dem virtuellen Desktop ablaufen anstatt auf dem Endgerät.

Für Unternehmen bedeutet Virtualisierung, das Management der Desktops zu zentralisieren und zu vereinfachen.[5] Der Aufwand bei herkömmlichen, traditionellen PC-Desktops ist umfassend und beinhaltet unter anderem die Installation, Wartung und Durchführung von Backups für jeden einzelnen PC. Durch die Zentralisierung der virtuellen Desktops auf Server ist es möglich, diesen Aufwand zu reduzieren. Dies zeigt sich zum Beispiel bei der Durchführung von Backups. Anstatt Daten von lokalen Systemen zu speichern, können diese zentral vom Rechenzentrum gesichert werden.[6] Auch andere Administrationsaufgaben lassen sich einheitlich und zeitsparend auf dem zentralen Server erledigen.[7]

Ein weiterer Vorteil der Virtualisierung gegenüber dem traditionellen Modell ist die Reduktion der Kosten. Dies bezieht sich auf Energie- sowie Hardwarekosten. Ein herkömmlicher PC fungiert als ganze Einheit mit einem eigenen Betriebssystem und eigenen Applikationen, die häufig jedoch nicht benötigt werden und unnötig Energie verschwenden. Durch die Virtualisierung werden nur die Ressourcen vom Server zur Verfügung gestellt, die der Anwender tatsächlich benötigt. Dadurch wird weniger Energie verwendet, weshalb sich Virtualisierung

[4] Enthalten in: o. V. (2014b)
[5] Vgl. Liebisch, D. (2010), S. 73
[6] Ebd., S. 76 ff
[7] Vgl. Lampe, F. (2010b), S. 93

als eine umweltfreundliche Möglichkeit herausstellt. Die Hardwarekosten werden durch einen längeren Hardware-Lebenszyklen ebenfalls reduziert, da die Hardware seltener ersetzt werden muss. Sollten die Kapazitäten im Laufe des Zyklus sich als ungenügend erweisen, sind Erweiterungen der maximalen Last auf Seiten des Servers mit wesentlich niedrigeren Aufwänden und Kosten verbunden, als dies bei herkömmlichen PCs der Fall wäre.

Zudem wird die Datenintegrität verbessert, da Daten zentral im Rechenzentrum gespeichert werden, anstatt Lokal auf dem PC. Zum einen wird dies durch die erleichterten Backups realisiert, zum anderen durch den Schutz der Daten selbst bei Verlust der Hardware. Des Weiteren gibt es weniger Softwarekonflikte, da weniger Programme auf dem Endgerät gespeichert sind.

Insgesamt ist festzustellen, dass im Vergleich zum herkömmlichen PC-Desktop die Virtualisierung größere Flexibilität[8] sowie Verwaltungs-, Sicherheits- und Kostenvorteile erlaubt.[9]

2.1.1.2 ThinClient

Für VDI-Umgebungen werden als Endgeräte traditionell ThinClients verwendet.[10] Da das Endgerät in der VDI-Umgebung deutlich weniger Rechenleistung erbringen muss als ein herkömmlicher PC, bieten sich ThinClients als Endgerät gut an. Sie sind klein, effizient, energiesparend[11] und besitzen keine Festplatte. Durch den Zugriff auf einen virtuellen Desktop bieten sie allerdings eine grafische Oberfläche und können mit einer Mouse bedient werden, sodass sich das Handling kaum von einem herkömmlichen PC-Desktop unterscheidet.[12] Abbildung 2 verdeutlicht, dass ThinClients nur für die Präsentation und für die Benutzungsschnittstelle verantwortlich sind. Die Anwendungslogik und die lokale Datenhaltung erfolgt auf dem Server, weshalb sie stark von diesem abhängig sind.

[8] Vgl. Redondo Gil, C. u. a. (2014), S. 37 f.
[9] Vgl. Greiner, W. (2010), S. 13
[10] Vgl. Liu, X./Sheng, W./Wang, J. (2011), S. 411
[11] Vgl. Greiner, W. (2010), S. 13
[12] Vgl. Lampe, F. (2010a), S. 93

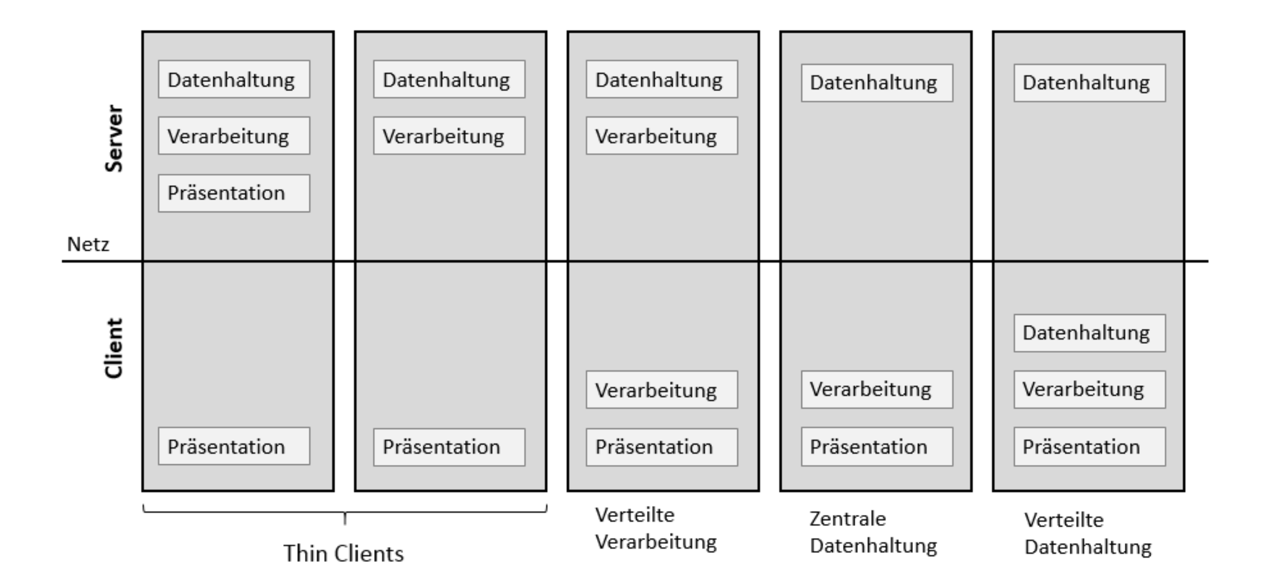


Abb. 2: Schaubild verschiedener Client-Server-Modelle[13]

Die Server sowie die Netzbelastung werden beim Einsatz von ThinClients stärker belastet als beim Einsatz von anderen Modellen, wie in Abbildung 2 gezeigt wird. Diese bezeichnet man auch als RichClients bzw. FatClients, die das Pendant zu ThinClients sind.[14]

Anwender können mit ThinClients auf virtuelle Desktops zugreifen. Dazu müssen sie sich zunächst Authentifizieren. Anschließend werden sie einem virtuellen Desktop zugewiesen oder es wird ein neuer initiiert.[15]

ThinClients zeichnen sich durch ihre hohe Zuverlässigkeit aus. Gründe dafür sind die einfache Hardware und die Begrenzungen hinsichtlich der lokalen Software und der Manipulationsmöglichkeiten. Diese stellen bei herkömmlichen PCs eine häufige Ausfallursache dar. Zudem verzichten sie komplett auf Lüfter sowie mechanische Laufwerke und Festplatten.

Mithilfe intelligenter Software- und Implementierungstools lassen sich ThinClients zentral oder von einem beliebigen Punkt im Netzwerk verwalten. Supportfahrten und lokale PC-Administrationen werden somit nicht mehr benötigt. Somit können Roll-outs von ThinClients sehr schnell durchgeführt werden. Dabei können ThinClients für den jeweiligen Arbeitsplatz vorkonfiguriert werden. Muss ein Gerät ausgetauscht werden, ist dies unproblematisch und kann ebenfalls sehr schnell durchgeführt werden.

Der lange Lebenszyklus von sechs bis acht Jahre machen ThinClients als Geräte für Unternehmen besonders attraktiv, denn er ist doppelt so lange wie der von herkömmlichen PCs. Zudem liegt der Stromverbrauch 50% unter dem eines PCs.[16]

[13] Mit Änderungen entnommen aus: Abts, D./Mülder, W. (2013), S. 139
[14] Abts, D./Mülder (W., 2013), S. 139
[15] Vgl. Lampe, F. (2010a), S. 102

2.1.2 Open Source (– Produkte)

Die Open Source Initiative (OSI) wurde 1998 gegründet und zeichnet sich mit dem Erfolg von Programmen aus, deren Quellcode für jeden frei zugänglich ist. Dabei spricht man von Open Source Software (OSS). Das Logo der OSI wird in Abbildung 3 gezeigt.



Abb. 3: Open Source Initiative Logo[17]

Das besondere Merkmal von OSS ist die Open Source Lizenz. Diese darf die Nutzung der Software nicht einschränken. Somit ist es möglich, die Software beliebig oft zu verwenden, ohne dafür eine Gebühr zahlen zu müssen. Deshalb wird OSS oft als kostengünstige Alternative zu proprietärer, kommerzieller Software gesehen. Es ist allerdings nicht zu unterschätzen, dass OSS programmiert, installiert und gewartet werden muss und diese Tätigkeiten in der Regel Kosten verursachen.

Es ist zudem zu beachten, dass individualisierte OSS, für deren Entwicklung eine Gebühr bezahlt werden kann, wiederum frei zur Verfügung stehen und der Quellcode offengelegt werden muss. Außerdem dürfen keine Einschränkungen bezüglich Nutzergruppen oder Einsatzfelder gemacht werden.

Die Lizenz muss technologieneutral sein, was bedeutet, dass sie keine spezielle Technologie, Software oder Schnittstelle voraussetzen darf. Dies erlaubt Unternehmen, unabhängig von Softwareherstellern zu sein, was bei kommerzieller Software häufig nicht der Fall ist.[18]

2.1.3 Kommerzielle ThinClient (– Produkte)

Der deutsche Marktführer unter den kommerziellen ThinClient Lösungen ist seit 2006 die deutsche IGEL Technology GmbH und einer der fünf größten Hersteller weltweit. IGEL ist im direkten Wettbewerb mit Wyse Technology (Dell), Hewlett-Packard und Fujitsu. Im Jahr 1997

[16] Vgl. Lampe, F. (2010b), S. 93
[17] Enthalten in: o. V. (2015)
[18] Vgl. Zehetmaier, J. (2011), S. 21 ff

wurde der erste moderne ThinClient produziert. IGEL gehört zu der Melchers-Gruppe und ist Mitglied von BITKOM. [19]

Das Produktportfolio richtet sich nach Linux- und Microsoft Windows-basierten Terminals mit jeweils unterschiedlichen Bauformen: Desktop-ThinClients, in LCD-Bildschirme integrierte Geräte und Zero Clients. Dazu vertreibt IGEL die Fernverwaltungssoftware IGEL Universal Management Suite (UMS) und die Software ThinClient Desktop Converter (UDC), welche ThinClients und ältere Windows PCs (XP), mit dem Betriebssystem IGEL Linux ausstattet. Ohne die IGEL UDC Software sind XP-Rechner nicht mehr weiter betriebsbereit, da Microsoft seit Anfang 2014 Sicherheits-Updates, Aktualisierungen und technischen Support für Windows XP eingestellt hat. [20]

IGEL Universal Management Suite (UMS) hingegen administriert ThinClients im Unternehmensnetz bequem, kostengünstig, ortsunabhängig, gruppenbasiert und sicher. Die Serviceleistung des UMS Produktportfolios umfasst ein Roll-out Gerätetausch. ThinClients werden dabei per Post versendet, lassen sich einfach anschließen und sich automatisch über das Netzwerk konfigurieren. Eine profilbasierte Administration ist ebenso gegeben, sodass Gruppen- und Einzelprofile per Drag und Drop zugewiesen werden können. Sichere Verbindungen, zertifikatsbasierte Anmeldung über SSL-Verschlüsselung, sowie remote, fail-safe, automatisierbare und bandbreitenoptimierte Firmware Updates, als auch Integrierte Tools, wie zum Beispiel Asset-Management zur automatischen Erfassung der Geräteinformation, sind ebenso Teil des IGEL UMS Lieferumfangs. Optional lassen sich Sicherheitslösungen, um Daten und Übertragungen zu schützen, an den Lieferumfang anpassen. Die IGEL Smartcard zur sicheren Zwei-Faktor-Authentifizierung, sowie USB-Token oder biometrische Geräte, Single Sign-on zur Verkürzung der Anmeldezeiten, Session-Roaming mit flexibler Wahl des ThinClient-Arbeitsplatzes, integrierter VPN-Client (Cisco, NCP), IEEE 802.1X zertifikatsbasierte Authentifizierung von ThinClients am Netzwerk-Switch, als auch Diebstalschutz, wie Kensington-Anschlüsse, sind IGEL Sicherheitslösungen, die in Anspruch genommen werden können. Die Vorteile des IGEL ThinClients wurden im oberen Teil dieser Arbeit bereits aufgezeigt. Durch den Service der IGEL Technology GmbH lassen sich jedoch noch weitere Vorteile aufzeigen. Zum einen ist zusätzlich eine lizenzkostenfreie Fernadministration mit inbegriffen. DVI- und VGA- Anschlüsse über Dualview Betrieb, sowie Single-Sign-on und schnelle Benutzerwechsel sind ebenso vorteilhafte Merkmale, genauso wie ein einfacher und sicherer IT-Support und User Help Desk für guten Kundenservice sorgen.[21]

[19] Vgl. IGEL Technology GmbH (2015)
[20] Vgl. Wurm, M. (2013)
[21] Vgl. IGEL Technology GmbH (2012)

Ein vorbildliches Beispiel liefert die Borussia VfL 1900 Mönchengladbach GmbH, dessen Wirtschaftlichkeit, Verfügbarkeit und Sicherheit der Stadion-IT-Umgebung mit der Einführung von IGEL ThinClient Arbeitsplätzen verbessert wurde. Windows NT-basierte PC-Arbeitsplätze wurden unter Citrix in eine Server basierte Computing-Umgebung innerhalb geringstem Konfigurations- und Administrationsaufwand eingebunden und schrittweise durch Igel ThinClients ersetzt. Smartcard-Anmeldungen wurden zur Verfügung gestellt, sodass nun eine eindeutige Zuordnung der Verkäufe an den Kassenterminals zu den Verkäufern garantiert und somit für mehr Transparenz im Ticketverkauf gesorgt werden kann. Da nun alle Prozesse der Borussia VfL 1900 Mönchengladbach GmbH auf einem einzigen Server durchgeführt werden, kommt es zu erheblich kostengünstigerem Aufwand für Wartungs- und Support-Kosten und führt zu Einsparungen von über 60.000 Euro. [22]

2.1.4 Virtual Private Network (VPN)

Ein Virtual Privat Network (VPN) ist ein logisches privates Netzwerk, das das Betreiben einer sicheren, virtuellen direkten Punkt-zu-Punkt-Verbindung zwischen zwei Stationen über ein öffentliches Netzwerk ermöglicht. Es garantiert einen Kommunikationsaufbau zwischen zwei Netzwerken über ein unsicheres Netzwerk, wie beispielsweise das Internet. [23]

Dabei ist es wichtig, dass die von Kommunikationspartnern ausgetauschten Daten und Informationen entsprechend verschlüsselt werden (Vertraulichkeit). Eine Authentifikation des VPNs wird durchgeführt, sodass nur autorisierte Benutzer Zugriff erhalten können und der VPN-Tunnel zum Daten- und Informationsaustausch ermöglicht wird (Authentizität). Außerdem muss sichergestellt werden, dass die Daten von Dritten nicht verändert werden können (Integrität). Daher sind die drei essentiellen Voraussetzungen des VPN-Verbindungsaufbaus Vertraulichkeit, Authentizität und Integrität der übertragenen Daten. Diese Voraussetzungen werden an jedem VPN-Endpunkt (Router oder Gateway) überprüft.

VPNs sind essentiell, um den Zugriff auf das Intranet von Institutionen zu ermöglichen, da die im Intranet vorhandenen Daten und Informationen durch die Firmen Firewall gesichert sind. Ein VPN hingegen ermöglicht durch die Authentifikation trotz Firewall den Zugriff auf das Firmen-Intranet. [24]

Eine gesicherte Datenübertragung über unsichere Netzwerke wird durch den Einsatz von Tunneling-Protokollen (IPSec, PPTP) garantiert, da so eine verschlüsselte Verbindung aufgebaut wird. Der Tunnel ist eine logische Verbindung zwischen zwei Endpunkten wie VPN-

[22] Vgl. IGEL Technology GmbH (2007)
[23] Vgl. Kappes, M. (2013), S. 197
[24] Ebd., S. 198 ff.

Clients, -Servern und -Gateways. Abbildung 4 zeigt den Aufbau der VPN-Verbindung von VPN-Server über den VPN-Tunnel bis hin zum VPN Client.

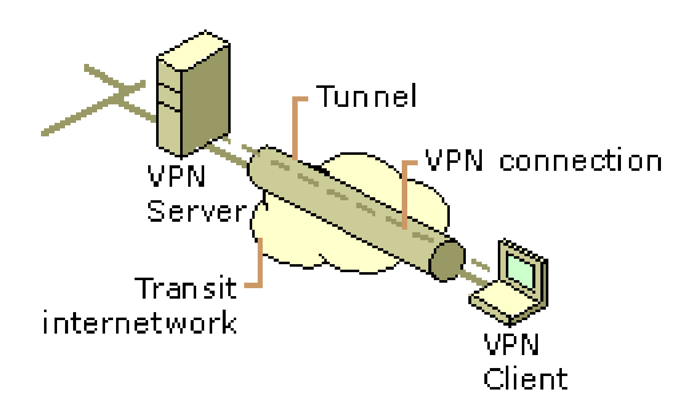


Abb. 4: Virtual Private Network Connection[25]

Es gibt hauptsächlich drei verschiedene VPN-Typen:

End-to-Site-VPN ist der klassische VPN-Verbindungsaufbau, um einen Remote-Access zu ermöglichen. Dabei ist das Ziel, dass externe Mitarbeiter, wie Heimatsarbeitsplätze oder mobile Benutzer, genauso arbeiten, als befänden sie sich im Netzwerk des Unternehmens. Eine logische Verbindung wird somit durch den VPN-Tunnel hergestellt. Dieser erschließt sich vom öffentlichen Netzwerk des externen Mitarbeiters (Host) zum entfernten lokalen Firmen-Netzwerk. Dieser Schritt wird in Abbildung 5 verdeutlicht.[26]



Abb. 5: End-to-Site-VPN Verbindung[27]

Site-to-Site-VPN, auch Branch-Office-VPN genannt, ermöglicht die Zusammenschaltung mehrerer lokaler Netzwerke, Außenstellen oder Niederlassungen, zu einem virtuellen Netzwerk über das öffentliche Netz. Es bietet sich aus Kostengründen an, statt einer physikalischen Festverbindung, bevorzugt die Internet-Verbindung zu benutzen. So können zwei oder mehr Netzwerke über einen VPN-Tunnel zusammengeschaltet werden (LAN-zu-LAN-Kopplung). Eine weitere Site-to-Site Variante ist das Extranet-VPN. Dies ermöglicht Dienste

[25] Enthalten in: o. V. (2001)
[26] Vgl. Schnabel, P. (2014)
[27] Enthalten in: Schnabel, P. (2014)

von fremden Unternehmen in das eigene Netzwerk zu integrieren und somit einen Zugriff von mehreren Firmen-Intranet-Netzwerken zu gewährleisten.[28]



Abb. 6: Site-to-Site-VPN Verbindung[29]

End-to-End-VPN, auch Host-to-Host- oder Remote-Desktop-VPN, ist der Verbindungsaufbau zwischen zwei oder mehreren Clients aus einem entfernten Netzwerk. Der VPN-Tunnel erschließt sich zwischen zwei Hosts. Ein direkter Verbindungsaufbau ist allerdings nicht möglich, da in der Regel eine Station zwischengeschaltet werden muss. Für jeden Client muss jeweils ein Gateway zur Verfügung gestellt werden, um die beiden Verbindungen zusammenschalten zu können.[30]



Abb. 7: End-to-End-VPN Verbindung[31]

2.1.5 Best practices

In vielen Unternehmen bremst die vorhandene IT-Infrastruktur das Kerngeschäft. Vor allem in der Versicherungs-Branche ist es notwendig, durch den Wettbewerbsdruck (viele Akquisitionen, ausländische Wettbewerber kommen hinzu) eine effiziente und langfristig kosteneinsparende IT-Landschaft aufzubauen. Virtualisierung mit ThinClients ist die Lösung für das Problem. Viele Versicherungsunternehmen transformieren ihre IT-Arbeitsplätze zu einer ThinClient IT-Infrastruktur. [32]

Das Frauenhofer-Institut für Umwelt-, Sicherheit- und Energietechnik belegt, dass durch Einsparungen bei den Support- , Strom- und Lizenzkosten bis zu 70% Kosteneinsparungen realisiert werden können. Grund dafür ist, dass häufig nur die Kosten für Lizenzen günstiger Terminalserver anfallen.[33]

[28] Vgl. Schnabel, P. (2014)
[29] Enthalten in: Schnabel, P. (2014)
[30] Vgl. Schnabel, P. (2014)
[31] Enthalten in: Schnabel, P. (2014)
[32] Vgl. IGEL Technology GmbH (2010)
[33] Vgl. IGEL Technology GmbH (2010)

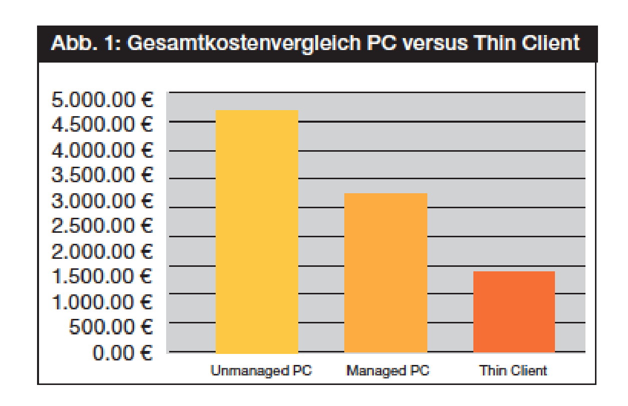


Abb. 8: PC vs. ThinClient - Wirtschaftsbetrachtung[34]

Aber nicht nur die Kosteneinsparungen sind vorteilhaft, sondern auch die erhöhte Sicherheit durch die zentrale Datenspeicherung und die hohe Verfügbarkeit geschäftskritischer Anwendungen. Durch die Zentralisierung von Anwendungen und die Standardisierung sowie Optimierung der IT bleibt mehr Zeit für strategische Aufgaben.

Openthinclient ist die herstellerunabhängige, lizenzfreie Lösung zur Bereitstellung hunderter von Arbeitsplätzen. Ihr Referenz-Profil deckt einige Branchen ab, darunter diverse Industrien, Gesundheitswesen und öffentliche Auftragsgeber.

Ein Paradebeispiel für Versicherungen, bei denen Openthinclient im Einsatz ist, ist die Oberösterreichische Versicherung AG aus Linz. Ihr Kundenportfolio erstreckt sich über 400.000 Kunden, was sie zum Marktführer für Schaden-Unfall und Leben in Oberösterreich auszeichnet. In Zusammenarbeit mit der Openthinclient GmbH arbeiten verschiedene Hardwarekomponente einfach und flexibel mit der Openthinclient Software zusammen. Die Oberösterreichische Versicherung hat insgesamt 50 Clients im Einsatz, die sich über die Citrix XenApp Farm verbinden lassen. Auch das Universitätsklinikum in Essen bietet in Kooperation mit der Openthinclient GmbH von den rund 6.000 PC-Arbeitsplätzen etwa 500 ThinClients an.[35]

Der deutsche Marktführer von kommerziellen ThinClient Lösungen, IGEL Technology GmbH, hat in Kooperation mit vielen Versicherungen eine ThinClient Umgebung geschaffen.

Die Basler Versicherung ermöglicht 400 angestellten den Zugriff über ThinClients auf Microsoft Standardapplikationen, Lotus Notes, das Internet und auf das Hostsystem. So wird eine zentrale Bereitstellung von Anwendungen über wartungsarme, fernadministrierte Endgeräte sichergestellt und Supportkosten werden dauerhaft gesenkt.

[34] Enthalten in: IGEL Technology GmbH (2010)
[35] Vgl. openthinclient GmbH (2014g)

Die Versicherung Lampe & Schwartze Bremen hat durch die IGEL ThinClient Lösung die ausgelagerte IT wieder zurück in den Betrieb geholt. Die Investition in ein neues Rechenzentrum konnte durch den Einsatz von ThinClients teilweise kompensiert werden. Nun arbeiten 150 Angestellte an Thin Clients, sodass ein Betriebskostenersparnis von 30% entsteht. Der Geschäftsführer bestätigte, dass die Lösung eine erhöhte Sicherheit durch die zentrale Datenspeicherung bietet, sie flexibel einsetzbar und sehr effizient ist, da sich schnell und einfach Gruppenprofile über das Netzwerk einrichten lassen.[36]

2.2 Methodisches Vorgehen: Erarbeiten eines Kriterienkatalogs zur Bewertung von VDI-Umgebungen

Im Folgenden wird erklärt, wie bei der Marktstudie vorgegangen werden soll. Dabei wird zunächst der Kriterienkatalog beschrieben, der dabei hilft, die Vergleiche verschiedener Thin-Client – Softwareprodukte durchzuführen. Die einzelnen Kriterien werden erläutert und beschrieben, um anschließend priorisiert und bewertet werden zu können. Es folgt eine abschließende Evaluation der Produkte mit dem Ergebnis, welches Produkt die Kriterien am besten erfüllt und sich für eine Empfehlung für den Kunden, in diesem Fall der .Versicherung, eignet.

2.2.1 Rechtfertigung und Beschreibung eines Kriterienkatalogs

Für die Durchführung einer Marktstudie ist es erforderlich, mehrere Produkte miteinander zu vergleichen. Um dieses Vorhaben zu begründen und systematisch zu erfüllen, ist es wichtig, die Anforderungen des Kunden anhand von Kriterien festzuhalten, zu priorisieren und zu bewerten. Somit kann letztendlich das geeignetste Produkt identifiziert werden. Ein Kriterienkatalog dient als Hilfestellung dafür und wird aus diesem Grund in dieser Arbeit verwendet. Zudem ist der Kriterienkatalog ein Ergebnis dieser Arbeit und ist im Rahmen der Aufgabenstellung des Projekts erwünscht.

In dem Kriterienkatalog werden alle Kriterien, die mit dem Kunden festgelegt werden, aufgelistet und priorisiert. Dies dient der anschließenden Bewertung. Anhand eines Bewertungsmodells, das im Folgenden beschrieben wird, kann das beste Produkt anhand des Vergleichs identifiziert werden. Die Identifikation wird somit deutlich effizienter und effektiver gestaltet. Das Ergebnis basiert auf einem klaren, systematischen Modell und findet somit seine Begründung.

[36] Vgl. IGEL Technology GmbH (2012)

2.2.2 Auflistung der Kriterien

Im Folgenden werden die Kriterien, die in dem Kriterienkatalog aufgelistet werden, vorgestellt. Der Kunde hat diese Kriterien vorgeschlagen und an das Projektteam kommuniziert. Sie beziehen sich auf die VDI-Umgebung und die dazugehörigen ThinClients.

2.2.2.1 Funktionsumfang

Der Funktionsumfang umfasst in dem Kontext der VDI-Umgebung die folgenden Unterpunkte:

- Verwaltungskonsole: Die Verwaltungskonsole dient als ein Verwaltungs- und Implementierungstool und hilft bei dem zentralen Management der virtuellen Desktops auf dem Server. Mit deren Hilfe können Administrationsaufgaben und andere Supporttätigkeiten zentral gesteuert werden, ohne einen Termin vor Ort zu benötigen. Für die Verwaltungskonsole stellt sich die entscheidende Frage, welche Funktionen sie umsetzen kann und in welchem Maße dies erfolgt. Zudem müssen deren Installationsvoraussetzungen Beachtung finden, um zu überprüfen, ob die Implementierung realisiert werden kann.

- Citrix-Fähigkeit: Die Citrix-Fähigkeit bezieht sich auf die ThinClients und ob Citrix auf diesen laufen kann. Dieses Kriterium sagt aus, ob die ThinClients in die VDI-Umgebung integriert werden können.

- VPN- Fähigkeit: Die VPN-Fähigkeit sagt aus, ob sich ein ThinClient remote durch ein Virtual Private Network (VPN) mit einem Server und somit zum virtuellen Desktop verbinden kann. Somit kann eine verschlüsselte und sichere Verbindung hergestellt werden, selbst wenn sich der Anwender nicht am Arbeitsplatz befindet.

2.2.2.2 Verbreitung

Die Verbreitung gibt an, im welchem Maße ein Produkt bereits in Unternehmen verwendet wird und wie bekannt es ist. Dies hat Auswirkungen auf die Kompatibilität und eventuellen Schnittstellen zu Hard- und Software. Dies ist ausschlaggebend für die Funktion und den Einsatz des Produkts.

2.2.2.3 Reife

Dieses Kriterium zeigt, zu welchem Grad die Reife und die Etablierung eines Produkts vorangeschritten sind. Dies lässt sich feststellen mit der Anzahl der Updates.

2.2.2.4 Lizenzmodell

Da in dieser Arbeit Open Source Produkte betrachtet werden, basiert das Lizenzmodell auf den OSI Lizenzen, die bereits im Grundlagenteil vorgestellt wurden. Somit sollte das Lizenzmodell bei allen Produkten ähnlich sein. Dennoch sollte ein genaues Augenmerk darauf

gelegt werden, um Verstöße gegen Lizenzrechtlinien und Urheberschutzbestimmungen zu vermeiden.

2.2.2.5 Implementierung

Die Einfachheit der Implementierung lässt sich anhand des Installiervorgangs festmachen. Dies sagt aus, ob die Realisierung der ThinClients mit geringem oder hohem Aufwand verbunden ist.

2.2.2.6 Gebühr

Die Gebühr und somit verbundene Kosten sind ein kritischer Punkt und müssen bei einem Vergleich ebenfalls in Betracht gezogen werden. Wünschenswert ist ein kostengünstiges Modell, allerdings müssen die Kosten ins Verhältnis zu den Leistungen gestellt werden. Es muss dabei zwischen Gebühren unterscheiden, die lediglich einmal anfallen und jenen, die regelmäßig gezahlt werden müssen.

2.2.3 Methodik der Bewertung

2.2.3.1 Priorisierung der Kriterien

Im Folgenden wird erklärt, wie die Kriterien priorisiert werden. Das Projektteam hat sich dafür ein Vorgehen überlegt, dass sich an der ABC-Analyse orientiert, um die Datenmenge der Kriterien übersichtlich zu strukturieren.37 Neben den Kategorisierungen „A“, „B“ und „C“ gibt es zusätzlich die „KO“ – Kategorie, die eine Abwandlung und Ergänzung zur ABC-Analyse darstellt. Die Kategorien werden im Folgenden beschrieben.

- **KO:** Zwingend Erforderliche Kriterien, die von höchster Priorität sind. Erfüllt ein Produkt nicht dieses Kriterien, ist es für den Vergleich zu eliminieren, da eine Kernanforderung nicht erfüllt wird.
- **A:** Diese A-Kriterien sind sehr wichtig und von hoher Bedeutung für den Vergleich der Produkte.
- **B:** B-Kriterien sind minder wichtig, deren Erfüllbarkeit ist dennoch von Interesse für den Kunden.
- **C:** Bei C-Kriterien handelt es sich lediglich um „nice to haves“.

Die Kriterien werden in die vier Kategorien eingeteilt und somit priorisiert. Für die Einteilung der Kriterien in die Kategorien gibt es weder ein spezielles Vorgehen noch bestimmte Richtlinien. Die Einteilung geschieht nach dem Ermessen des Projektteams. Dabei sind die Vorgaben des Kunden ausschlaggebend, denn sie geben an, auf welche Kriterien der Fokus

[37] Vgl. Schawel, C./Billing, F. (2004), S. 13 ff.

gelegt werden muss. Im Folgenden wird beschrieben, welcher Kategorie die Kriterien zugeteilt werden:

Verwaltungskonsole

Die Verwaltungskonsole stellt für den Kunden in diesem Projekt eine wichtige Funktion dar, die unbedingt als Teil der VDI-Umgebung bestehen sollte. Aus diesem Grund wurde dieses Kriterium als sehr wichtig kategorisiert und muss dementsprechend als ein Kriterium der Kategorie A priorisiert werden.

Citrix-Fähigkeit

Um als ThinClient – Endgerät das Betriebssystem aufrufen und somit überhaupt eine Funktion ausführen zu können, ist dieses Kriterium von sehr hoher Bedeutung. Aus diesem Grund muss es als KO-Kriterium priorisiert werden. Ist ein Produkt nicht Citrix-fähig wird es eliminiert und nicht weiter bei der Marktstudie untersucht.

VPN - Fähigkeit

Die Fähigkeit, eine VPN-Verbindung mit dem Server aufzustellen wird als unwichtig angesehen. Grund für diese Priorisierung sind die Schilderungen des Kunden, nach denen diese Funktion nur wenige Male im Jahr von wenigen Mitarbeitern genutzt wird. Deshalb ist es möglich, diesem Kriterium wenig Bedeutung beizumessen und es der C-Kategorie zuzuordnen.

Verbreitung

Dieses Kriterium sagt aus, wie bekannt ein Produkt ist und wie oft es bereits in Unternehmen verwendet wird. Da es somit eine Auswirkung auf die Kompatibilität und auf Schnittstellen hat, beeinflusst es die Funktion und den Einsatz des Produkts. Dennoch ist es möglich, dass ein neues Produkt, das noch nicht sehr verbreitet ist, die beste Lösung für ein Unternehmen darstellen kann. Anhand dieser Überlegungen ist dieses Kriterium als wichtig einzustufen. Die Einstufung in die A-Kategorie wurde jedoch als zu hoch empfunden, weshalb das Produkt mit seiner Entwicklung für die B-Kategorie geeigneter ist.

Reife

Die Reife eines Produkts ist sehr wichtig, denn es sagt aus, zu welchem Grad es bereits vorangeschritten ist. Mit einer hohen Anzahl von Updates ist davon auszugehen, dass es sich

bereits um ein fortgeschrittenes Produkt handelt, bei dem schwerwiegende Fehler und Bugs unwahrscheinlich sind. Dies hat starke Auswirkungen auf die Funktion und Zuverlässigkeit eines Produkts, sodass die Reife als Kriterium der A-Kategorie zuzuteilen ist.

Lizenzmodell

Da in diesem Projekt lediglich Open Source – Produkte betrachtet werden, sind für die Marktstudie nur Produkte zulässig, die den Open Source Lizenzen entsprechen. Ist dies nicht der Fall, ist das Produkt unzulässig und muss eliminiert werden. Deshalb entspricht dieses Kriterium der KO-Kategorie.

Implementierung

Die Einfachheit der Implementierung ist wichtig, um Aussagen über die Realisierbarkeit zu machen. Aus diesem Grund ist dieses Kriterium wichtig. Allerdings kann es vorkommen, dass Produkte aufgrund ihrer Komplexität einen gewissen Grad an Aufwand haben. Dennoch kann das Produkt dennoch oder gerade aus diesem Grund das geeignetste sein, weshalb die B-Kategorisierung bei diesem Kriterium sinnvoll ist.

Gebühr (regelmäßige Zahlungen)

Gebühren sind grundsätzlich Kosten und daher ein wichtiger Faktor beim Vergleichen mehrerer Produkte. Daher ist vor allem wegen der regelmäßigen Zahlung ein hoher Aufwand beim Lizenzmanagement und bei der Buchhaltung verbunden. Die hohe Gewichtung ist sinnvoll, weshalb sich die A-Kategorisierung in diesem Fall eignet.

Gebühr (einmalige Zahlung)

Im Vergleich zur regelmäßigen Zahlung stellt die einmalige Zalhung deutlich weniger Aufwand dar. Aus diesem Grund ist die Gewichtung dieses Kriteriums weniger wichtig, sodass sich die C-Kategorisierung anbietet.

Die Priorisierungen anhand der zugeteilten Kategorie werden in Tabelle 1 nochmals übersichtlich aufgezeigt.

Kriterium	Kategorie
Funktionsumfang: Citrix-Fähigkeit	KO

Lizenzmodell	KO
Funktionsumfang: Verwaltungskonsole	A
Reife	A
Gebühr (regelmäßige Zahlungen)	A
Verbreitung	B
Einfachheit der Implementierung	B
Funktionsumfang: VPN-Fähigkeit	C
Gebühr (einmalige Zahlung)	C

Tab. 1: Übersicht der Kategorisierung der Kriterien[38]

Die Kategorien werden für die Berechnung der abschließenden Evaluation benötigt.

2.2.3.2 Bewertung der Kriterien

Die Bewertung der Kriterien zeigt an, in welchem Maße die Kriterien von den jeweiligen Produkten erfüllt werden. Dabei muss eine Punktzahl vergeben werden, die zwischen 0 und 10 ist, wobei 1 das niedrigste und 10 das höchste Ergebnis ist.

Verwaltungskonsole (inkl. Funktionsumfang und Installationsvoraussetzung)

Die Verwaltungskonsole ist ein wichtiges Kriterium, um festzustellen, ob einerseits die Administration benutzerfreundlich aufgesetzt wurde und andererseits wie einfach der Zugriff auf die Verwaltungskonsole ist und welche Möglichkeiten der Benutzer hat. Dabei wurden Verwaltungskonsolen sehr hoch bewertet, die im Browser über graphische, Web-basierte Oberflächen dargestellt werden (Bewertung > 5). Hingegen Verwaltungskonsolen, die lokal administriert werden müssen, bedeuten für den Benutzer einen ungeheuer großen Aufwand und werden somit mit einer geringen Punktzahl bewertet (Bewertung < 5).

Citrix-Fähigkeit

Grundsätzlich kann ein Produkt im Produkt-Lieferumfang Citrix bedingungslos unterstützen (Bewertung 10) oder auch gar nicht zur Verfügung stellen (Bewertung 0). Allerdings gibt es auch die Möglichkeit, dass es Produkte gibt, die nicht von Anfang an Citrix-fähig sind, jedoch als Zusatz Citrix zum Lieferumfang hinzugefügt werden kann. Bei der Bewertung dieser Ausnahmen ist es davon abhängig, wie die Erfahrungsberichte ausgefallen sind, ob das Customizing umständlich ist (Bewertung < 5), oder im Gegenteil einfach und zufriedenstellend (Bewertung > 5).

VPN-Fähigkeit

[38] Eigene Darstellung

Ist ein Produkt VPN-fähig, so wird es mit 10 Punkten bewertet. Wenn dem nicht so ist, wird die geringste Punktzahl vergeben. Für die Ausnahme, dass beispielsweise ein Anbieter die VPN-Fähigkeit in Zukunft anbieten wird, da derzeit die Entwicklung noch im Gange ist, so fällt die Bewertung gut oder schlecht aus, je nach Wahrscheinlichkeit, ob dieser Zusatz bereits in naher Zukunft erwerblich ist.

Verbreitung

Das Kriterium Verbreitung wird abhängig von der Anwenderzahl bewertet. Je mehr Anwender, desto besser fällt die Bewertung aus. Zu einer guten Bewertung führt allerdings auch, wenn Firmenreferenzen eine ähnliche Infrastruktur wie der Kunde aufweisen können (Bewertung > 5). Kann das Produkt keine Referenzen aufweisen, so wird dem Kriterium Verbreitung die geringste Punktzahl zugeordnet (Bewertung < 5).

Reife

Der Reifegrad spielt ebenso eine große Rolle. Ist das Produkt auf dem aktuellen Stand, bietet sinnvolle, regelmäßige Updates an, so fällt die Bewertung sehr gut aus (Bewertung > 5). Wobei die Häufigkeit von Updates keine große Rolle spielt, sondern einerseits vor allem auch die Kontinuität des Produktes. Es wäre schließlich suboptimal, wenn das Produkt nach der Einführung vom Markt verschwinden würde (Bewertung < 5). Andererseits ist es ebenso von großer Bedeutung, dass der Anbieter keine kommerziellen Absichten hat.

Lizenzmodell

Das Kriterium Lizenzmodell lässt sich bewerten, indem teure Lizenzmodelle gering (Bewertung < 5) und günstige Lizenzmodelle gut (Bewertung > 5) bewertet werden. Ist allerdings absehbar, dass das Lizenzmodell teilweise kommerzialisiert ist, so werden minimale Abstriche bei der Bewertung anfallen.

Einfachheit der Implementierung

Die Einfachheit des Implementierungsprozesses ist davon abhängig, inwiefern der Benutzer von der jeweiligen Implementierungs-Dokumentation geführt wird. Muss sich der Benutzer mühselig durch eine Textanleitung durchkämpfen, so fällt die Bewertung nicht so gut aus (Bewertung < 5) wie ein Implementierungsprozess, der auf einem graphischen Interface basiert und somit die Administration und Installation ohne weitere Umstände durchgeführt werden können. (Bewertung > 5)

Gebühr (einmalige/regelmäßige Zahlung)

Je angemessener der Preis, desto besser fällt die Bewertung aus. Stimmt das Preisleistungsverhältnis von einem teureren Produkt aufgrund von Zusatzfunktionen kann auch dabei eine gute Bewertung entstehen.

2.2.3.3 Abschließende Evaluation

Für die abschließende Evaluation werden die Bewertungspunktzahlen vom vorherigen Schritt benötigt. Diese werden mit einem Faktor multipliziert, der sich auf die Kategorisierung bezieht.

Dabei ist es wichtig, dass die wichtigen Kriterien, mit der Kategorie A den Faktor 3 zugeordnet bekommen. Denn je höher der End-Wert der Bewertung desto besser schneidet das Produkt ab. Somit wird der Kategorie B der Faktor 2 zugeteilt und der weniger wichtigen Kategorie C der Faktor 1.

KO Kriterien, wie beispielsweise die Citrix Fähigkeit oder Lizenzmodell, bekommen keinen Faktor, da sie entweder existieren oder nicht. Somit stellt sich nur die Frage, besitzt das jeweilige Produkt dieses Kriterium, ja oder nein. Ist die Antwort nein, so wird das Produkt von der Bewertung ausgeschlossen.

Kategorie	*Faktor*
KO	-
A	3
B	2
C	1

Tab. 2: Faktorzuteilung für die Berechnung der Evaluation[39]

Das Beispiel der Abbildung 7 verdeutlicht die Evaluation der Ergebnisse. Dabei werden 2 Produkte mit einander verglichen und auf jeweils zwei Kriterien getestet. Die Kategorie und die Bewertung des jeweiligen Kriteriums spielen dabei die entscheidende Rolle.

Erläuterung

Der Reifegrad des Produkts 1 schneidet mit 9 von 10 Punkten sehr gut ab, wohingegen die Verbreitung nur 5 Punkte erzielt hat. Allerdings ist der Reifegrad durch die Kategorie A entscheidender, als die Verbreitung, mit der Kategorie B. Um den Wert des Produktes zu berechnen rechnet man 3 (Kategorie A) x 9 (Bewertung Reife Prod 1) + 2 (Kategorie B) x 5 (Bewertung Verbreitung Prod 1) = (3 x 9) + (2 x 5) = 27 + 10 = 37.

[39] Eigene Darstellung

Somit besitzt **Produkt 1** den **Wert 37**.

Simultan wird der gleiche Schritt mit Produkt 2 durchgeführt. 3 (Kategorie A) x 5 (Bewertung Reife Prod 2) + 2 (Kategorie B) x 9 (Bewertung Verbreitung Prod 2) = (3 x 5) + (2 x 9) = 15 + 18 = 33.

Produkt 2 besitzt also den **Wert 33**.

Dieser Wert ist zunächst nur ein Zwischenergebnis. Denn das Zwischenergebnis wird noch durch einen weiteren Wert **dividiert**, der abhängig von den Kategorien der Kriterien ist. Die Kriterien Reife A und Verbreitung B haben den **Wert 5**, da Wert von A = 3 und Wert von B = 2.

Also gilt 2 + 3 = **5**.

Um das finale Ergebnis des Kriteriums zu erhalten dividiert man die Werte 37 und 33 jeweils durch 5.

Produkt 1: 37/5 = 7,4 Produkt 2: 33/5 = 6,6

Somit hat Produkt 1 besser abgeschnitten als Produkt 2.

Beispiel:

Produkt	Kriterium	Kategorie (A, B, C)	Bewertung (0-10)	Berechnung	Wert
Prod. 1	Reife	A	9	(3 x 9)	27
	Verbreitung	B	5	(2 x 5)	10
					37
Prod. 2	Reife	A	5	(3 x 5)	15
	Verbreitung	B	9	(2 x 9)	18
					33

Prod. 1: Prod. 2:

$$\frac{37}{(3+2)} = 7{,}4 \quad > \quad \frac{33}{(3+2)} = 6{,}6$$

Tab. 3: Rechenbeispiel[40]

[40] Eigene Darstellung

3 Praktisches Vorgehen

3.1 Bewertung

Entlang der im Rahmen des Kriterienkatalogs definierten Bewertungsrubriken werden im Folgenden die Thinclient-Betriebssysteme openthinclient und ThinStation evaluiert. Anschließend werden ein Punktwert bezüglich der Erfüllung des jeweiligen Anforderungsmerkmales vergeben und das Gesamtergebnis ermittelt, entsprechend der im vorangegangenen Abschnitt festgelegten Skala und Gewichtung.

3.1.1 Funktionsumfang

Verwaltungskonsole

Bereits die grundlegende Administrierbarkeit unterscheidet beide Betriebssysteme signifikant. Während die kommerzialisierte Lösung openthinclient eine benutzerfreundliche, webbasierte und ohne zusätzliche Software erreichbare Managementschnittstelle bietet, erfolgt die Administration von ThinStation primär terminalgestützt und mit Hilfe von Client-seitig gespeicherten Konfigurations- und Hostdateien.[41] Für diese scheint keine mitgelieferte zentrale Wartungsmöglichkeit zu existieren, wodurch der Verwaltungsaufwand bei Änderungen der Netzwerkinfrastruktur oder Konfiguration erheblich höher einzustufen ist.[42] Auch das Hinzufügen neuer Clients gestaltet sich mit openthinclient einfacher, da dies mit wenigen Klicks auf der grafischen Verwaltungsoberfläche durchgeführt werden kann.

Aufgrund der mangelnden Administrationskonsole erhält ThinStation eine sehr geringe Wertung von 2, welche lediglich die vorhandenen, eher umständlichen Konfigurationsmöglichkeiten wertschätzt. Dementgegen wurde an openthinclient aufgrund der besseren Administrierbarkeit, der hohen Flexibilität und dem mächtigeren Funktionsumfang eine 8 vergeben.

Citrix-fähig

Ein Kriterium, welches eine kritische Anforderung darstellt ist die Kompatibilität mit einer Citrix VDI Infrastruktur, wie diese bereits beim Kunden vorhanden ist. Hierbei erreichen aufgrund der Vorauswahl nach KO-Kriterien beide Anwendungen die volle Punktzahl.

Openthinclient unterstützt zwar in seiner Grundinstallation nicht das Citrix-ICA-Protokoll, jedoch ist diese Funktionalität nachträglich durch hinzufügbare Anwendungen erweiterbar.[43] Diese Lösung wurde bereits erfolgreich bei einem Referenzkunden aus dem Versicherungs-

[41] Vgl. openthinclient GmbH, (2014d), S. 1
[42] Vgl. Cupp Jr., D. A. (2014c)
[43] Vgl. openthinclient GmbH (o. J.)

sektor mit Citrix XenApp Infrastruktur implementiert.[44] ThinStation dementgegen unterstützt das Citrix ICA Protokoll bereits grundsätzlich.[45]

VPN-fähig

Die Implementierbarkeit des Betriebssystems innerhalb einer nur über VPN mit dem Firmennetzwerk verbundenen Umgebung hängt grundsätzlich von der Möglichkeit ab, das Betriebssystem lokal auf dem Clienten permanent speichern zu können. Jedoch unterstützt openthinclient nur PXE-Boot, wodurch keine dauerhaften lokalen Kopien möglich sind. Allerdings befindet sich ein Erweiterungspaket gerade in der Fundraising-Phase und wird voraussichtlich in mittlerer Zukunft verfügbar sein, womit dann per Administratoranweisung auch lokale Replikate erzeugt werden können, deren Konfiguration sich bei jedem Verbinden mit dem Netzwerk mit den Servereinstellungen synchronisiert.[46] Das lokal gespeicherte Betriebssystem kann hierbei auf nahezu jedem beliebigen Medium abgelegt werden wie zum Beispiel einem USB-Stick. Da die Entwicklung der Erweiterung noch nicht sicher ist und dementsprechend auch nicht testbar, erhält openthinclient eine Bewertung von nur 3.

ThinStation unterstützt grundsätzlich ohne Notwendigkeit von Erweiterungen das lokale Speichern des Systems auf einer Festplatte oder einem USB-Speichermedium, wodurch es 10 Punkte erhält.[47]

3.1.2 Verbreitung

Anwenderzahl

Innerhalb der rund sechsjährigen Firmengeschichte der openthinclient GmbH konnte diese eine solide Anzahl an Referenzen Referenzen aufbauen. Hierbei handelt es sich zum Teil sogar um überregional agierende Unternehmen wie den Keramikhersteller Duravit oder das deutsche rote Kreuz.[48] Hinzu kommt die Österreichische Versicherung AG, welche eine Citrix-gestützte Infrastruktur umgesetzt hat mit etwa 50 Clients.[49] Aufgrund dessen erhält openthinclient für dieses Kriterium 9 von 10 Punkten.

Bei ThinStation ist die tatsächliche Verbreitung schwierig einzuschätzen, insbesondere im gewerblichen Umfeld, da keine Artikel zu Referenzen vorliegen und sich lediglich einige Er-

[44] Vgl. openthinclient GmbH (2014g)
[45] Vgl. Cupp Jr., D. A. (2014c)
[46] Vgl. openthinclient GmbH (2014b)
[47] Vgl. Cupp Jr., D. A. (2014c)
[48] Vgl. openthinclient GmbH (2014g)
[49] Vgl. openthinclient GmbH (2014c)

wähnungen der Software in einigen IT-Magazinen finden.[50] Hierdurch wurde ThinStation mit 5 bewertet.

3.1.3 Reife

Update-Häufigkeit

Die Entwicklung von openthinclient und ThinStation erfolgen primär auf GitHub. Beide Softwarelösungen verfügen dort über stetig neue Beträge. Bei openthinclient fallen diese Aktivitäten jedoch in größerer Zahl und Frequenz an, was auf die kommerzielle Motivation der Entwickler zurückzuführen sein kann. Zusätzlich können Kunden selbst Erweiterungen vorschlagen, welche anschließend geprüft werden auf ihre Komplexität. Danach wird ein Crowdfunding-Aufruf auf der openthinclient-Website gestartet, um die notwendigen Entwicklungskosten zu finanzieren.[51] Alles in allem erhält deswegen diese Lösung 10 Punkte. ThinStation kann lediglich eine ebenfalls hohe GitHub-Aktivität vorweisen, wodurch ein kleiner Abschlag auf 8 Punkte zustande kommt.[52]

Etabliertheit

Openthinclient kann auf etwa 6 Jahre durchgehende Entwicklung und Anpassung durch ein profit-orientiertes Unternehmen zurückblicken, wodurch die Software über ein deutlich höheres Entwicklungsstadium verfügt und somit im Vergleich als relativ reif betrachtet werden kann.[53] Zudem existiert die von mehreren IT-Unternehmen getragene OpenThinclientAlliance, welche sich ebenfalls für die Weiterentwicklung und Verbreitung der Software engagiert.[54] Somit erhält openthinclient 10 Punkte.

ThinStation hingegen wird lediglich primär von einer Privatperson getragen, wodurch der Entwicklungsgrad der Software sich von der anderen Lösung unterscheidet und die Kontinuität der Entwicklung als unsicherer betrachtet werden kann.[55]

3.1.4 Einsatz in großen Unternehmen

Bei der Verbreitung bei großen Unternehmen erlangt openthinclient ebenfalls eine gute Bewertung von 9 Punkten, da sogar Referenzen aus dem Versicherungsbereich mit Citrix XenApp Infrastruktur vorhanden sind und generell eine umfangreiche Liste an großen Referenz-

[50] Vgl. o. V. (2014a)
[51] Vgl. GitHub Inc. (2015a)
[52] Vgl. GitHub Inc. (2015b)
[53] Vgl. openthinclient GmbH (2014d)
[54] Vgl. Open Thin Client Alliance (2011)
[55] Vgl. Cupp Jr., D. A., (o. J.)

kunden vorliegt.[56] ThinStation kann bei diesem Kriterium keine Punkte erhalten, da sich keine Referenzkunden finden lassen.

3.1.5 Lizenzmodell

Freie Zugänglichkeit

Das Lizenzmodel bei beiden Betriebssystemen basiert fast ausschließlich auf GPL2. Einzig openthinclient verfügt über Erweiterungen, welche mit Hilfe anderer Open Source Lizenzen lizensiert sind, um eine ausschließlich entgeltliche Nutzung dieser Komponenten zu erwirken.[57, 58] Dadurch erhält openthinclient einen leichten Abschlag auf 9 Punkte, während ThinStation 10 erhält.

3.1.6 Gebühren

Einmalig

Auch wenn openthinclient an sich kostenfrei ist, würde der Einsatz von openthinclient in einer Citrix-Umgebung einmalige Lizenzkosten verursachen, da Komponenten, welche insbesondere für den gewerblichen Einsatz der Software bestimmt sind, wie zum Beispiel eine Citrix Receiver Applikation, nur gegen ein Gebühr von 9,50 € pro ThinClient und 105 € pro Serverinstanz angeboten werden.[59] Das führt zu lediglich 3 Punkten in dieser Wertung. ThinStation ist allerdings komplett kostenfrei und erhält dafür 10 Punkte.[60]

Regelmäßig

Regelmäßige Gebühren gibt es bei beiden Lösungen nicht, weshalb beide die volle Punktzahl von 10 erhalten.[61, 62]

3.1.7 Einfachheit der Implementierung

Installationsvorgang

Der Installationsvorgang der Systeme verhält sich analog zur Verwaltungskonsole. Dies bedeutet, dass openthinclient über ein Setupprogramm mit einer grafischen Benutzeroberfläche verfügt, welches einen automatisch durch alle notwendigen Schritte der Installation führt und

[56] Vgl. openthinclient GmbH (2014g)
[57] Vgl. openthinclient GmbH (2014e)
[58] Vgl. Cupp Jr., D. A. (2014b)
[59] Vgl. openthinclient GmbH (o. J.)
[60] Vgl. Cupp Jr., D. A. (2014c)
[61] Vgl. openthinclient GmbH (2014d)
[62] Vgl. Cupp Jr., D. A. (2014c)

somit die Installation relativ einfach und nicht sehr fehleranfällig gestaltet.[63] Hierdurch erhält openthinclient 8 Punkte.

ThinStation verfügt im Gegensatz dazu nur über einen terminal-basierten und sehr manuellen Installationsprozess, für welchen nur eine textbasierte Anleitung vorliegt.[64] Dadurch gestaltet sich die Einrichtung deutlich komplizierter und Fehleranfälliger. Ebenfalls finden sich zu ThinStation diverse Berichte über Schwierigkeiten beim Einrichten der Software, wodurch sie nur 3 Punkte für dieses Kriterium erhält.

[63] Vgl. openthinclient GmbH (2014h)
[64] Vgl. Cupp Jr., D. A. (2014a)

3.2 Gesamtbewertung

Abschließend lässt sich feststellen, dass openthinclient in allen Kriterien außer der Gebühren deutlich besser im Vergleich mit ThinStation abschneidet. Die folgende Tabelle zeigt die anhand der Gewichtungen zusammengefassten Ergebnisse der Bewertung.

Kriterium	Gewichtung	openthinclient	Thin Station
Funktionsumfang	A	6,75	4
Verbreitung	B	9	5
Reife	A-	9,125	6,4
Lizenzmodell	KO	9	10
Gebühren	B	8,25	10
Einfachheit der Implementierung	B	8	3

Tab. 4: Übersicht – Gesamtbewertung[65]

Bildeten man den gewichteten Durchschnitt aller Kriterien mit der bisher verwendeten Gewichtungsmethodik, dann erhält man ein deutliches Ergebnis zugunsten openthinclient. Diese Software erhält damit eine Gesamtwertung von rund 8,2 Punkten, während ThinStation 5,6 erhält.

Die aggregierten Ergebnisse aller Kriterien sind in der folgenden Grafik aufbereitet.

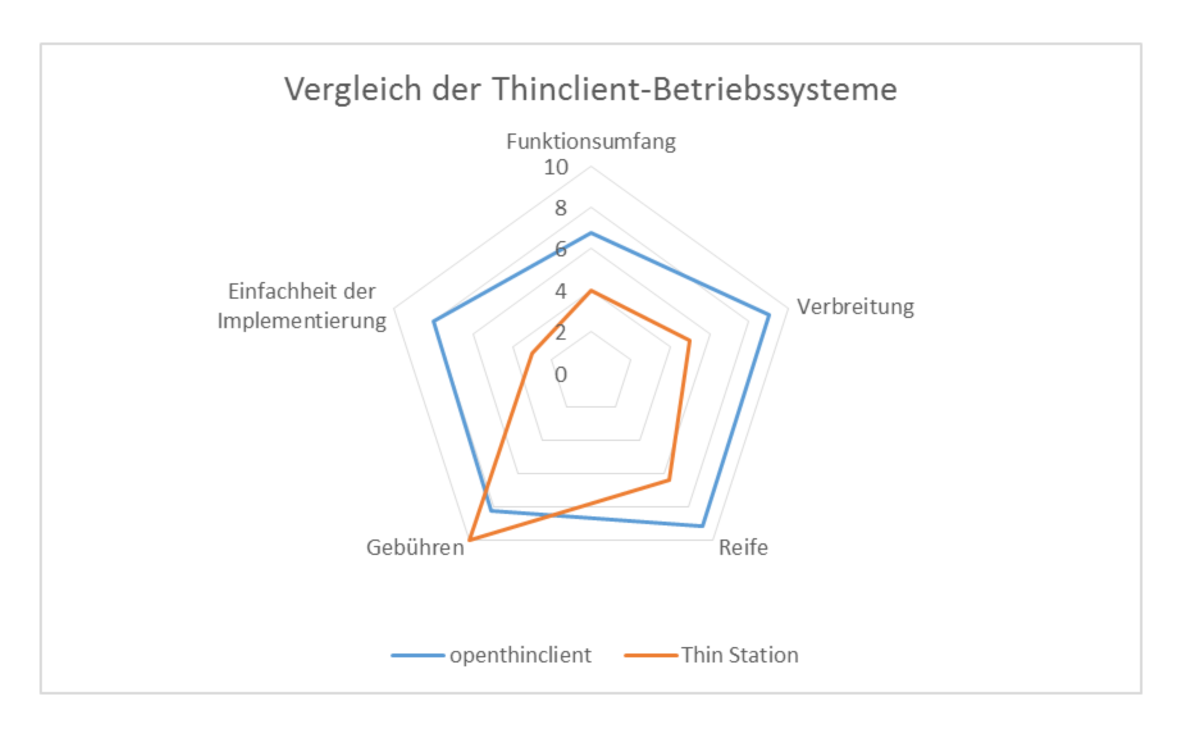


Abb. 9: Auswertung der Bewertungen als Spinnennetz-Graph[66]

[65] Eigene Darstellung
[66] Eigene Darstellung

Aufgrund der nur mittleren Gewichtung des Kriteriums der Gebühren und des deutlichen Vorsprungs von openthinclient in sämtlichen Kategorien kann diese Software für den Einsatz bei dem Kunden .Versicherung vorgeschlagen werden, da sich dessen Anforderungen am besten von allen betrachteten Betriebssystemen erfüllt.

Kritische Würdigung

Das Bewertungsergebnis ist grundsätzlich von der Richtigkeit der Entwicklerangaben abhängig, da nicht alle Kriterien, wie zum Beispiel die Implementierung innerhalb einer VPN-Umgebung getestet werden konnten und somit weitgehend auf die offiziellen Angaben der Hersteller vertraut wird. Ebenfalls erhöht die nicht vollständige Verfügbarkeit von Informationen bezüglich ThinStation die Ungenauigkeit der Bewertung. Nichtsdestotrotz handelt es sich bei openthinclient um eine verbreitete und vielfach erprobte Software mit soliden Referenzen und Kritiken, weshalb diese nach bestem Wissen und Gewissen empfohlen wird.

4 Testumgebung

Abschließend soll eine Testumgebung verwendet werden, anhand derer die von openthinclient zur Verfügung gestellten Informationen zu technischen Fragestellungen des Kriterienkatalogs validiert werden können. Diese Testumgebung soll anhand eines Testszenarios erarbeitet werden, das eine rudimentäre openthinclient-Umgebung darstellt und im Folgenden entwickelt wird.

4.1 Testszenario

Eine Umgebung, die für den Einsatz von openthinclient geeignet sein soll, muss über einige Komponenten verfügen. Diese umfassen diverse ThinClients, die das Booten über ein Netzwerk via PXE/TFTP beherrschen, einen Server, auf dem die openthinclient Software installiert werden kann, sowie ein lokales Netzwerk (LAN), das diese Geräte vernetzt und über einen DHCP-Server verfügt. Der ebenfalls als Vorrausetzung gegebene Terminal Server ist für das vorliegende Testszenario irrelevant, da nur die Administration der ThinClients, nicht die Umsetzung einer Virtual Desktop Infrastructure evaluiert werden soll.[67] Der DHCP Server ist hierbei jedoch unerlässlich, da ein PXE-Boot nur unter Verwendung dieses Protokolls möglich ist.[68]

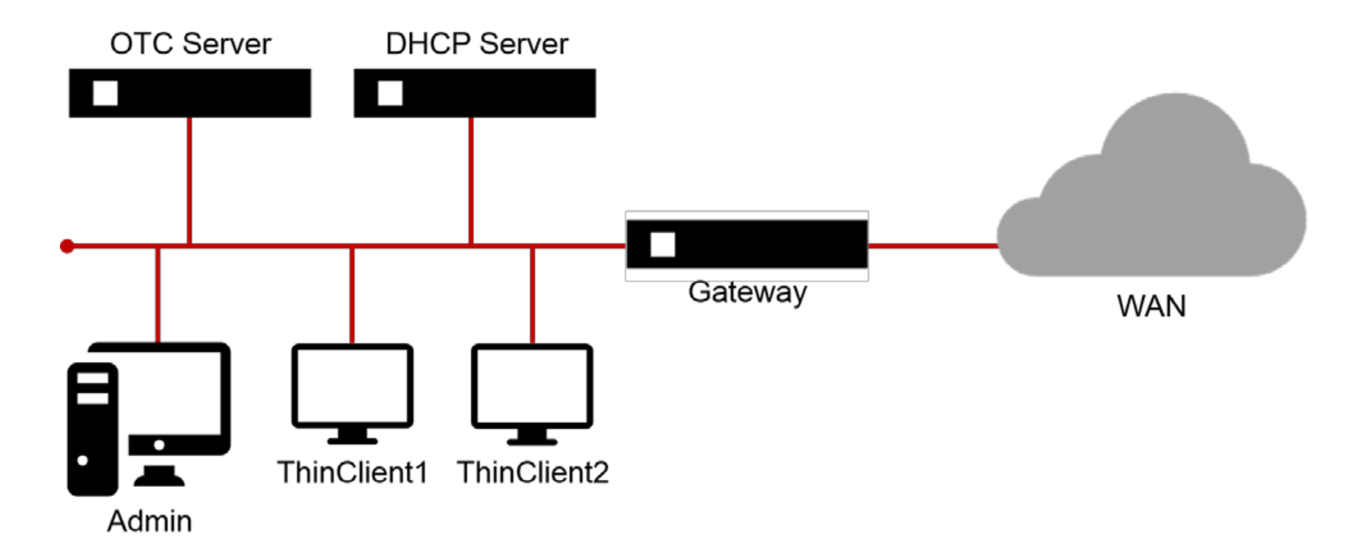


Abb. 10: Schematische Darstellung des Testszenarios[69]

[67] openthinclient GmbH (2015)
[68] Vgl Henry, M., (1999), S. 1
[69] Eigene Darstellung

Im Testszenario, auf dem die zu entwickelnde Testumgebung basieren wird, sollten demnach ThinClients, ein Server sowie ein DHCP Server vorhanden und über ein LAN verbunden sein, um den Mindestanforderungen der openthinclient-Lösung zu genügen. Wie in Abb. 10 zu sehen ist, lässt sich das Testszenario um zwei weitere Komponenten ergänzen. Zum einen ist es möglich, die Verwaltung der openthinclient-Umgebung statt direkt auf dem Server mithilfe eines weiteren Rechners im Netzwerk vorzunehmen. Da der Fokus des Testszenarios auf der Evaluierung der Administrierbarkeit der Umgebung liegt, sollte diese Möglichkeit integriert werden. Zum anderen ist die Anbindung des lokalen Netzwerkes über ein Gateway an ein Weitverkehrsnetz (WAN) zu sehen. Über diese Komponente können die anderen Geräte Verbindungen in das Internet aufbauen, was beispielsweise bei der Installation von Zusatzsoftware zu enormer Zeitersparnis führen kann. Die Abgrenzung des Testnetzwerks von dem Internetzugangsnetz mithilfe des Gateways ist enorm wichtig, da in ersterem ein DHCP Server betrieben wird, der nicht mit einem DHCP-Server im Zugangsnetz konkurrieren darf.

4.2 Virtualisierte Repräsentation

Da die Testumgebung in einer Form erstellt werden soll, die möglichst problemlos dem Kunden übergeben werden kann, erscheint eine Umsetzung mit dedizierter Hardware für die einzelnen Komponenten ungeeignet. Eine Alternative wäre, sämtliche Komponenten als virtuelle Maschinen zu erstellen. Dies würde zu einer einfachen Migration der Testumgebung verhelfen, da virtuelle Maschinen problemlos als Datei abgespeichert, versendet und in einer anderen Virtualisierungsumgebung erneut betrieben werden können. In Absprache mit dem Kunden wurde daher entschieden, das entwickelte Testszenario in eine virtualisierte Repräsentation zu überführen.[70]

4.2.1 Virtualisierungsumgebung

Um eine Testumgebung mit virtuellen Maschinen erstellen zu können, ist es zunächst notwendig, sich für eine mögliche Virtualisierungsumgebung zu entscheiden. Zu unterscheiden ist hierbei zum einen zwischen dem Hersteller der entsprechenden Umgebung, zum anderen ob es sich um einen Hypervisor handelt, der als eigenständiges Betriebssystem oder als Applikation auf einem PC installiert wird.

Um eine Entscheidung bezüglich des Herstellers der Virtualisierungstechnologie treffen zu können, soll deren Etabliertheit herangezogen werden. Zu diesem Zwecke soll das Gartner Magic Quadrant für x86 Server Virtualisierung aus dem Jahr 2014 betrachtet werden. Dieses visualisiert die Ergebnisse einer von Gartner durchgeführten Marktanalyse, die die Hersteller

[70] Vgl. Fritzsche, A. (2015)

von Virtualisierungsinfrastrukur anhand ihrer Vision und der Funktionsfähigkeit ihrer Produkte den Kategorien „Leaders“, „Challengers“, „Visionaries“ und „Niche Players“ zuordnet.[71]



Abb. 11: Gartner Magic Quadrant für x86 Server Virtualisierung[72]

Wie in Abb. 11 zu sehen ist, wird die Kategorie „Leader“ nur von VMware und Microsoft belegt, wobei VMware sowohl in der Vollständigkeit der Vision, als auch in der Funktionsfähigkeit seiner Produkte bessere Ergebnisse erzielt als Microsoft. Aus diesem Grunde fällt die Wahl des Herstellers auf VMware.

Schließlich gilt es noch zu entscheiden, welches Produkt von VMware eingesetzt werden soll. Im Wesentlichen lässt sich hier zwischen den für den Desktopgebrauch gedachten Player und Workstation sowie dem vSphere Hypervisor für Servervirtualisierung unterscheiden.[73] Der erstgenannte VMware Player lässt sich jedoch direkt ausschließen, da hier sowohl die Möglichkeit einer umfassenden Konfiguration der virtuellen Netzwerke, als auch eine Funktion zum Exportieren von virtuellen Maschinen nicht integriert ist.[74] Unter anderem im Bereich der Möglichkeiten zur Netzwerkkonfiguration, was für den Aufbau einer ThinClient-Umgebung enorm wichtig ist, zeigt sich, dass vSphere mehr Möglichkeiten bietet als

[71] Vgl. Gartner Inc. (2015)
[72] Enthalten in: Bittman, T./Margevicius, M./Dawson, P. (2014)
[73] Vgl. VM Ware Inc. (2015b)
[74] Vgl. VM Ware Inc. (2015c)

Workstation.[75] Da zudem das auf Konsolenbasis implementierte Betriebssystem des vSphere deutlich ressourcenschonender arbeitet als beispielsweise ein Windowssystem, auf dem VMware Workstation installiert ist, ist es sicherlich die beste Alternative, die Entscheidung zu Gunsten des VMware vSphere zu treffen. Die für den Testaufbau zur Verfügung stehende Hardware, ein HP ProLiant MicroServer mit 8GB Arbeitsspeicher, genügt hierbei den Installationsanforderungen des vSphere 5.5.[76]

4.2.2 Netzwerkkonfiguration

Um das lokale Netzwerk des Testszenarios darzustellen, genügt es, auf dem vSphere Hypervisor ein virtuelles Netzwerk einzurichten, dem alle virtuellen Maschinen innerhalb des lokalen Netzwerkes zugeordnet sind. Um dieses lokale Netzwerk betreiben zu können, insbesondere hinsichtlich des PXE Boot, ist es notwendig, einen DHCP-Server zu installieren. Um Ressourcen zu sparen soll hierbei jedoch von der Möglichkeit Gebrauch gemacht werden, diesen DHCP-Server gemeinsam mit dem Gateway für den Internetzugriff auf einer virtuellen Maschine zu implementieren.

Zu diesem Zweck muss ein Betriebssystem für die virtuelle Maschine ausgewählt werden, das einerseits den Betrieb als Gateway sowie des DHCP-Servers ermöglicht, andererseits aber auch ressourcensparend betrieben werden kann. Aus diesen Gründen eignet sich ein konsolenbasiertes Serverbetriebssystem. Möglichkeiten gibt es in diesem Bereich einige, aus Kostengründen soll jedoch auf ein Open Source Linux zurückgegriffen werden. Da sämtliche Linux-Serverlösungen die gegebenen Kriterien erfüllen, ist die endgültige Auswahl für das Ergebnis von geringer Auswirkung. Eine Möglichkeit, auf die in dieser Testumgebung zurückgegriffen werden soll, ist Ubuntu Server 14.04.

Die virtuelle Maschine, die als Gateway dienen soll, muss mit zwei virtuellen Netzwerkkarten ausgestattet werden, um ein Routing zwischen dem internen und dem externen Netzwerk betreiben zu können. Hierbei soll die erste an das nach außen gerichtete virtuelle Netzwerk, das zweite an das interne verknüpft werden.

Zunächst soll auf die Implementierung des DHCP Servers eingegangen werden. Dies ist mit der kostenlosen Lösung dnsmasq einfach umzusetzen. Die Konfigurationsdatei dieser Software muss für das vorliegende Szenario nur mit dem gewünschten DHCP-Adressbereich für das interne Netzwerk beschrieben werden. In diesem Fall soll das Netzwerk 192.168.1.0/24 mit dem DHCP-Range 192.168.1.100-192.168.1.199 verwendet werden.

[75] Vgl. VM Ware Inc. (2015a)
[76] Vgl. VM Ware Inc. (2014), S. 11

Der Betrieb als Gateway zwischen dem internen und dem externen Netzwerk erfordert zudem die Konfiguration der iptables des Ubuntu-Servers. Hier müssen Regeln für das Zulassen aller ausgehenden sowie bereits bestehenden Verbindungen erstellt werden. Zudem ist es notwendig, mithilfe der IP Masquerading-Funktion, Network Adress Translation zu betreiben.

4.2.3 Openthinclient Server

Für den Betrieb der openthinclient Lösung, ist es notwendig einen Server mit der entsprechenden Software auszustatten. Hierbei lässt sich zwischen der Installation einer Java-Anwendung auf einem existierenden Server oder der Verwendung einer fertig konfigurierten virtuellen Maschine entscheiden.[77] Da der openthinclient-Server aufgrund der integrierten DHCP-Proxy Funktion jedoch nicht auf derselben Maschine wie der DHCP-Server des Netzwerkes installiert werden darf[78] und daher der bereits existierende Server nicht in Frage kommt, bietet es sich an, die fertige virtuelle Maschine auf den vSphere aufzuspielen. Vor dem ersten Start sollte jedoch sichergestellt werden, dass die virtuelle Netzwerkkarte mit dem internen virtuellen Netzwerk verknüpft ist. Da auf diesen Server manuell zugegriffen werden soll, ist es sinnvoll eine statische IP Adresse (192.168.1.10) außerhalb des DHCP-Range zu vergeben.

4.2.4 Administration

Die Administration des openthinclient Servers erfolgt über eine Java Anwendung, die mithilfe eines Browsers geladen werden kann. Dies kann zwar auf dem Server selbst erfolgen, sollte jedoch aus Gründen der Handhabbarkeit auch vom PC des Systemadministrators aus funktionieren. Um dies darstellen zu können, soll ein Desktopbetriebssystem auf der bestehenden Umgebung aufgesetzt werden. Grundsätzlich sollten sich alle Betriebssysteme eignen, die über eine grafische Oberfläche, einen Browser und eine Java Runtime verfügen. Aus Kostengründen wird hier wieder auf die Open Source Lösung Ubuntu 14.04 zurückgegriffen. In diesem Fall jedoch soll die Desktop Variante aus den genannten Anforderungen verwendet werden. Die Einstellung der Netzwerkverbindung, die IP Adresse über DHCP zu beziehen, kann beibehalten werden.

4.2.5 ThinClients

Das Aufsetzen und Konfigurieren der virtuellen Maschinen zum Betrieb als ThinClients umfasst nur einen minimalen Aufwand. Es genügt, virtuelle Maschinen ohne Festplatte mit einer Netzwerkverbindung in das interne virtuelle Netzwerk zu erzeugen.

[77] Vgl. openthinclient GmbH, (2015)
[78] Vgl. openthinclient GmbH, (2014f)

4.2.6 Virtualisierte Testumgebung

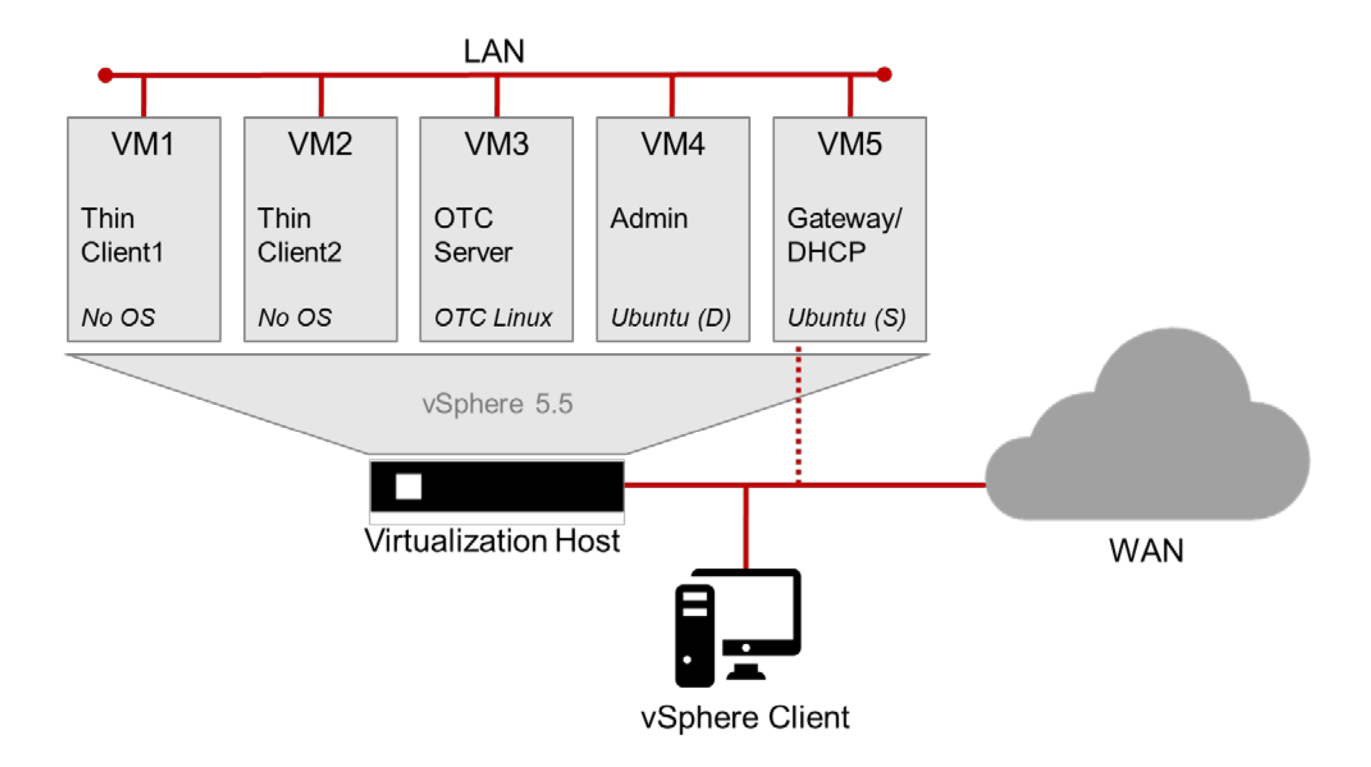


Abb. 12: virtualisierte Testumgebung[79]

Das Resultat der Überlegungen und vorgenommenen Konfigurationen der vorigen Abschnitte zeigt sich in Abb. 12. Fünf virtuelle Maschinen (zwei ThinClients, der openthinclient Server, der Admin sowie das Gateway mit DHCP Server) sind auf einem Host virtualisiert, der mit vSphere 5.5 betrieben wird. Dieser ist an ein Netzwerk angebunden das Zugriff auf ein Weitverkehrsnetz (z.B. Internet) hat. Um den virtuellen Maschinen ebenfalls Zugriff auf das WAN zu geben, dient die VM5 als Gateway. Schließlich fällt bei der Betrachtung auf, dass eine zusätzliche Komponente hinzugefügt wurde. Beim vSphere Client handelt es sich um einen PC, der die Software zur Administration und Konfiguration des vSphere Hypervisor sowie der virtuellen Maschinen installiert hat.

4.3 Validierung der Bewertung

Zu guter Letzt soll anhand der Testumgebung validiert werden, inwiefern die openthinclient-Lösung den Erwartungen gerecht wird, auf deren Basis die Bewertung erfolgte. Hierfür soll zunächst analysiert werden, welche Bewertungskriterien sich anhand des virtuell erstellten Testszenarios prüfen lassen.

[79] Eigene Darstellung

4.3.1 Auswahl überprüfbarer Bewertungskriterien

Innerhalb der sechs Bewertungskategorien des in Kapitel 2.2 entwickelten Kriterienkatalogs, eignen sich nur der Funktionsumfang sowie die Einfachheit der Implementierung für eine genauere Betrachtung bei der Überprüfung in der Testumgebung. Die restlichen Kategorien (Verbreitung, Reife, Lizenzmodell, Gebühren) stellen keine technischen Fragestellungen dar und sind daher nicht validierbar.

Um bei der Durchführung der Überprüfung ein chronologisches Vorgehen beizubehalten, soll zunächst auf die Einfachheit der Implementierung eingegangen werden. Dieser Schritt ist geprägt vom Installationsvorgang des openthinclient-Systems. Da in der Testumgebung die bereits vorkonfigurierte virtuelle Maschine verwendet wird, gilt es zu validieren, ob es tatsächlich genügt, diese in eine bestehende Umgebung zu importieren sowie die Netzwerkkonfiguration vorzunehmen, um die Implementierung abzuschließen.

Die Kategorie Funktionsumfang setzt sich zusammen aus den Kriterien Verwaltungskonsole, Citrix-Fähigkeit und VPN-Fähigkeit. Ersteres lässt sich innerhalb der Testumgebung überprüfen, indem getestet wird, inwiefern die Managementschnittstelle intuitiv gestartet und eingerichtet werden kann. Zudem soll hierbei die Schwierigkeit der Konfiguration neuer ThinClients evaluiert werden. Das Kriterium der Citrix-Fähigkeit lässt sich mangels einer bestehenden Citrix-VDI-Umgebung sowie fehlender Lizenzen für den Citrix Receiver bei openthinclient ebenso wenig prüfen wie die VPN-Fähigkeit, die am reinen PXE-Boot scheitert, wie in Kapitel 3.1.1 erläutert wurde.

Kriterium	Geeignet zur technischen Validierung
Funktionsumfang: Verwaltungskonsole	ja
Funktionsumfang: Citrix-fähig	ja (nicht durchführbar)
Funktionsumfang: VPN-fähig	ja (nicht durchführbar)
Verbreitung	-
Reife	-
Lizenzmodell	-
Gebühren	-
Einfachheit der Implementierung	ja

Tab. 5: Eignung der Kriterien zur technischen Validierung[80]

Wie in Tab. 5Tab. 1 zusammenfassend dargestellt, eignen sich demnach nur wenige Kriterien zur Überprüfung innerhalb der Testumgebung, wovon die Validierung nur für den Funktionsumfang der Verwaltungskonsole sowie die Einfachheit der Implementierung durchführbar ist. Dennoch stellt dieser Abschnitt einen wichtigen Beitrag zur Auswahl einer geeigneten

[80] Eigene Darstellung

Lösung dar, da eine Empfehlung auf möglichst umfangreichen Untersuchungen basieren sollte. Zudem kann eine funktionierende Testumgebung auch über die jetzigen Tests hinaus, in Zukunft für die Überprüfung von Erweiterungen des Systems verwendet werden und stellt somit eine wichtige Ergebniskomponente dar.

4.3.2 Einfachheit der Implementierung

Um die vorkonfigurierte virtuelle Appliance zu installieren, muss diese zunächst heruntergeladen werden. Das entsprechende zip-Archiv ist hierbei leicht über den Download-Bereich der openthinclient-Website zu finden.[81] Der Import auf den VMware vSphere erfolgt nach dem Entpacken bequem über den vSphere Client. In der abschließenden Konfiguration zum Import, genügt es, die virtuelle Netzwerkkarte („bridged“) dem vSwitch zuzuordnen, der innerhalb des Hypervisors das lokale Netzwerk repräsentiert.

Beim ersten Boot-Vorgang erhält die virtuelle Maschine eine dynamische IP-Adresse vom DHCP-Server. Nach dem Anmelden mit den vorkonfigurierten Zugangsdaten (User: openthinclient, Passwort: openthinclient), erfolgt die Einrichtung einer statischen IP-Adresse (192.168.1.10) in der Netzwerkkonfiguration, die aus der grafischen Benutzeroberfläche heraus gestartet werden kann, wie in Abb. 13 rot hervorgehoben ist. Der openthinclient-Server läuft bereits als Dienst im Hintergrund, kann jedoch nach Belieben oder Bedarf ebenfalls über die GUI (grüne Markierung in Abb. 13) neu gestartet werden.

[81] Vgl. openthinclient GmbH (2014a)

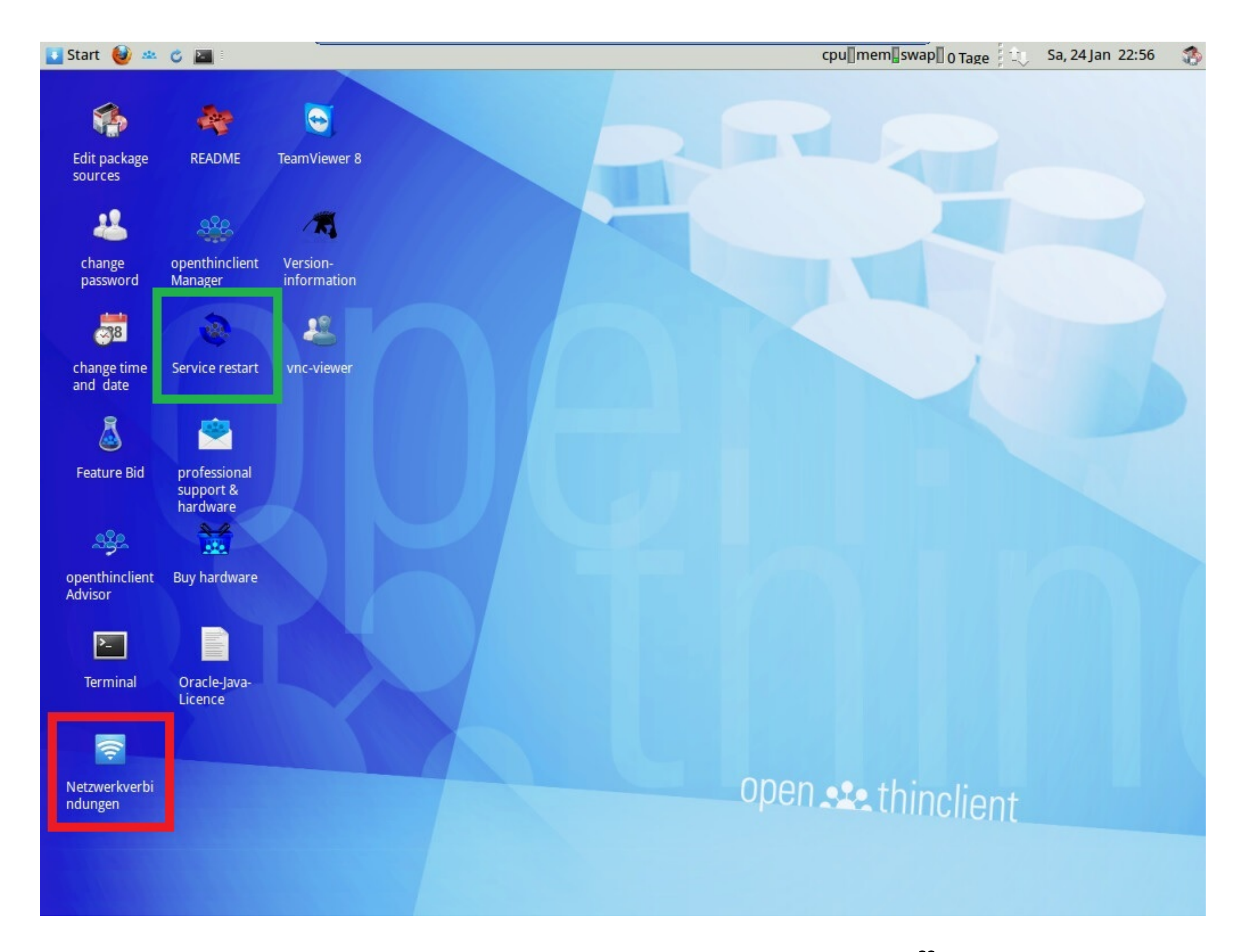


Abb. 13: Desktop der openthinclient-Virtual-Appliance[82]

Insgesamt zeigt sich die Installation der openthinclient-Virtual Appliance als benutzerfreundlich, intuitiv und einfach, weshalb an dieser Stelle die gute Beurteilung der Software hinsichtlich dieses Kriteriums bestätigt werden kann.

4.3.3 Verwaltungskonsole

Die Validierung des Starts und der Einrichtung der Verwaltungskonsole soll innerhalb der für die Administration des Systems vorgesehenen virtuellen Maschine mit Ubuntu 14.04 Desktop Version erfolgen. Eine Voraussetzung für das Öffnen der Verwaltungskonsole ist die Unterstützung von Java ab Version 1.6.[83] Nach erfolgreicher Installation der Java Laufzeitumgebung ist zum Start der Administrationsoberfläche von openthinclient der Aufruf der IP-Adresse des Servers (192.168.1.10) auf dem http-Proxy-Port (8080) in einem Browser notwendig. Die angezeigte Webseite führt über einen Klick auf „Start Manager" zum Download und Start des openthinclient Managers.

[82] Screenshot: eigene Darstellung
[83] Vgl. openthinclient GmbH (2014f)

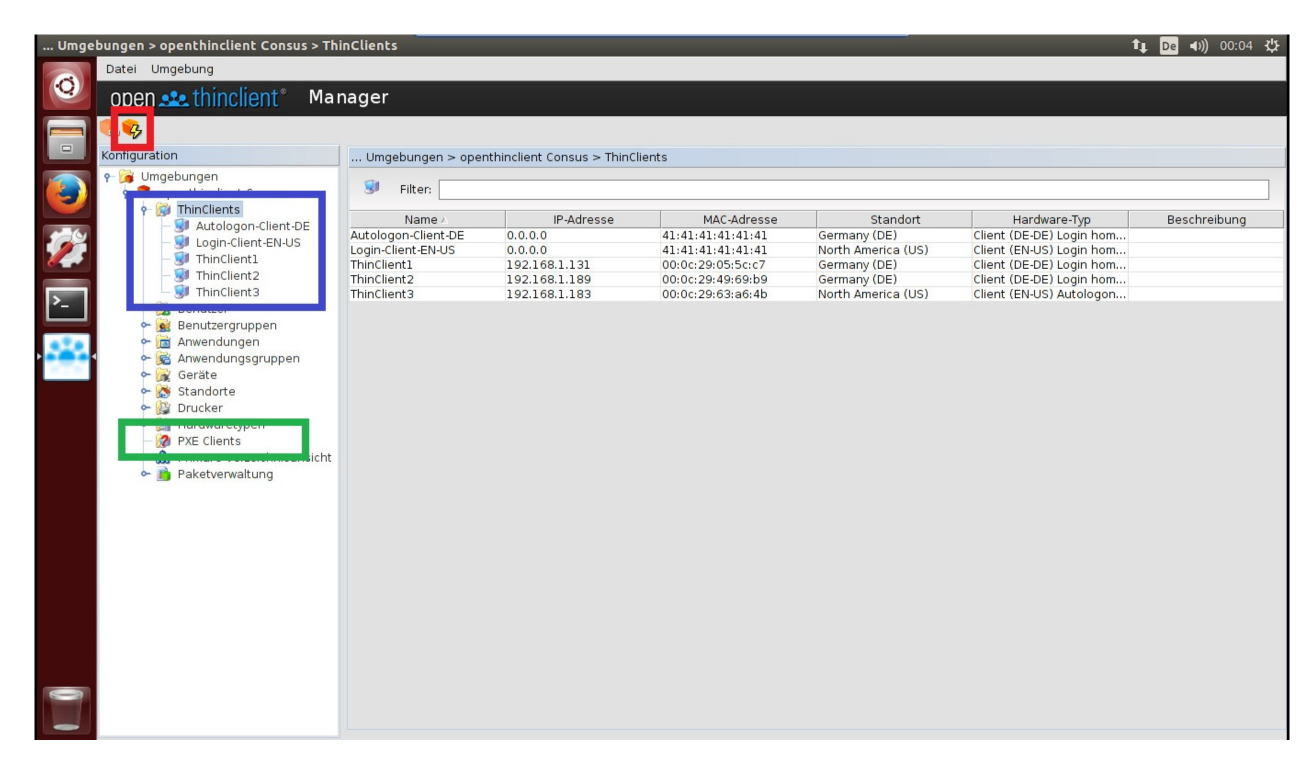


Abb. 14: Verwaltungskonsole openthinclient Manager[84]

Nach dem Öffnen der Applikation muss diese mit dem implementierten openthinclient-Server verbunden werden, um diesen verwalten zu können. Dieser Vorgang erfolgt über die „Mit Verwaltungsumgebung verbinden"-Schaltfläche, die im Screenshot in Abb. 14 rot hervorgehoben wurde. Im sich öffnenden Dialog genügt es, das vorkonfigurierte Passwort der Umgebung („0pen%TC") einzugeben, um die Verbindung herzustellen. Die IP-Adresse des Servers sowie der verwendete LDAP-Port (10389) sind ebenso voreingetragen, wie der Benutzername.[85] Nach erfolgreicher Einbindung der Umgebung, kann diese nun verwaltet werden. Auch nach einem Neustart der Software wird die bereits eingerichtete Verbindung wieder hergestellt.

Der Start und die Konfiguration der Verwaltungskonsole sind demnach, wie zu erwarten war, entsprechend der Installation des Servers relativ einfach, intuitiv und übersichtlich. Bleibt zu klären, inwieweit dies bei der Einrichtung von ThinClients beibehalten wird.

Grundsätzlich existieren zwei Möglichkeiten, einen neuen ThinClient in der Umgebung anzumelden. Einerseits kann mit einem Rechtsklick auf ThinClients (blaue Markierung in Abb. 14) ein neuer ThinClient mithilfe seiner MAC-Adresse registriert werden. Zum anderen registriert der openthinclient-Server jeglichen Versuch eines PXE-Boots im lokalen Netzwerk. Versucht ein noch nicht registrierter ThinClient einen solchen Boot über das Netzwerk, so

[84] Screenshot: eigene Darstellung
[85] Vgl. openthinclient GmbH (2014f)

wird er unter der Rubrik „PXE-Clients“ (grüne Markierung in Abb. 14) aufgeführt. Über das Kontextmenü dieses Eintrags lässt sich der entsprechende Client registrieren.

Bei einem PXE-Bootversuch eines Clients mit registrierter MAC-Adresse sollte dieser nun fähig sein, das ihm zugeordnete openthinclient-Betriebssystem zu beziehen und zu laden.[86] Leider war der Bootvorgang in der Testumgebung entgegen den Instruktionen des Herstellers erst erfolgreich, nachdem der openthinclient-Server als PXE-Boot-Server im DHCP-Dienst der Gateway-VM eingetragen wurde. Dieser Eintrag erfolgte mittels der Anweisung „dhcp-boot=/pxelinux.0,openthinclient-server,192.168.1.10“ in der Konfigurationsdatei des dnsmasq-Dienstes. Anschließend war das Booten der ThinClients jedoch problemlos möglich.

Abschließend betrachtet kann demnach auch die hohe Bewertung der Verwaltungskonsole als gerechtfertigt angesehen werden. Der Zugriff sowie die Einrichtung des openthinclient-Managers gestalteten sich völlig problemlos. Auch das Einrichten der ThinClients war ohne Schwierigkeiten zu bewältigen. Lediglich die zusätzlich notwendige Konfiguration des DHCP-Servers stellte ein unerwartetes Hindernis dar, welches jedoch einfach und effektiv überwunden werden konnte und dessen Lösung keinen wiederkehrenden Aufwand darstellt.

[86] Vgl. openthinclient GmbH (2014f)

5 Abschließende Betrachtung

5.1 Reflexion der Zielerreichung

Basierend auf den Ergebnissen der Marktstudie zur gegenwärtigen Marktlage der ThinClient-Betriebssysteme wurden zwei zur Evaluierung geeignete Kandidaten ermittelt. Um die ausgewählten Systeme zu bewerten, wurde ein Kriterienkatalog auf Basis gängiger Methoden zur Beurteilung von Open Source Software erstellt.

Im nächsten Schritt konnte unter Verwendung der Bewertungsmethodik des Kriterienkatalogs openthinclient als geeigneteres System ermittelt werden. Abschließend kann festgestellt werden, dass auf Basis des Kriterienkatalogs eine geeignete Softwarelösung ermittelt werden konnte, welche die erwarteten Anforderungsmerkmale erfüllt. Anhand der aufgebauten Testumgebung konnte dies schließlich bestätigt werden.

5.2 Ausblick

Aufgrund der stetigen Weiterentwicklung von openthinclient durch die dazugehörige Entwicklungsfirma, kann dieses als zukunftsweisendes und nachhaltiges Produkt angesehen werden. Auch ist die langfristige Versorgung mit Systemaktualisierungen und –erweiterungen aufgrund der Schlüsselrolle der Software im Produktmix der openthinclient GmbH gewährleistet. Zusätzlich stellen ThinClients eine ökologisch nachhaltige Form der Arbeitsplatzausstattung dar, da sie wenig Energie verbrauchen und wegen ihrer Softwareunabhängigkeit längere Lebenszyklen besitzen. Schließlich kann festgehalten werden, dass die openthinclient-Software eine vielversprechende Lösung für die Umsetzung einer Virtual Desktop Infrastructure bei der .Versicherung darstellt.

Anhang

Quellenverzeichnisse

Literaturverzeichnis

Abts, D. / Mülder, W. (2013): Grundkurs Wirtschaftsinformatik : Eine kompakte und praxisorientierte Einführung, 8. Aufl, Wiesbaden: Springer Vieweg

Greiner, W. (2010): Die Grünende IT : Wie die Computerindustrie das Energiesparen neu erfand, in: Frank Lampe (Hrsg.) Green-IT, Virtualisierung und Thin Clients : Mit neuen IT-Technologien Energieeffizienz erreichen die Umwelt schonen und Kosten sparen, 1. Aufl., Wiesbaden: Vieweg+Teubner Verlag, S. 3–16

Kappes, M. (2013): Netzwerk- und Datensicherheit : Eine praktische Einführung, 2.Aufl., Wiesbaden: Springer Vieweg

Lampe, F. (Hrsg.) (2010): Green-IT, Virtualisierung und Thin Clients : Mit neuen IT-Technologien Energieeffizienz erreichen die Umwelt schonen und Kosten sparen, 1. Aufl., Wiesbaden: Vieweg+Teubner Verlag

Lampe, F. (2010a): Thin Clients : Anwendungsvirtualisierung (SBC) oder Desktop-Virtualisierung?, in: Frank Lampe (Hrsg.) Green-IT, Virtualisierung und Thin Clients : Mit neuen IT-Technologien Energieeffizienz erreichen die Umwelt schonen und Kosten sparen, 1. Aufl., Wiesbaden: Vieweg+Teubner Verlag, S. 101–109

Lampe, F. (2010b): Thin Clients : Eine Einführung, in: Frank Lampe (Hrsg.) Green-IT, Virtualisierung und Thin Clients : Mit neuen IT-Technologien Energieeffizienz erreichen die Umwelt schonen und Kosten sparen, 1. Aufl., Wiesbaden: Vieweg+Teubner Verlag, S. 91–100

Liebisch, D. (2010): Desktop-Virtualisierung, in: Frank Lampe (Hrsg.) Green-IT, Virtualisierung und Thin Clients : Mit neuen IT-Technologien Energieeffizienz erreichen die Umwelt

schonen und Kosten sparen, 1. Aufl., Wiesbaden: Vieweg+Teubner Verlag, S. 71–90

Liu, X. / Sheng, W. / Wang, J. (2011): Design of Multimedia Computer Room Based on Virtual Desktop Infrastructure, in: Gang Shen/Xiong Huang (Hrsg.) Advanced research on computer science and information engineering : International conference, CSIE 2011, Zhengzhou, China, May 21-22 2011, proceedings, 152-153, Heidelberg: Springer, S. 410–414

Niemer, M. (2010): Stromsparen durch Virtualisierung, in: Frank Lampe (Hrsg.) Green-IT, Virtualisierung und Thin Clients : Mit neuen IT-Technologien Energieeffizienz erreichen die Umwelt schonen und Kosten sparen, 1. Aufl., Wiesbaden: Vieweg+Teubner Verlag, S. 57–69

Redondo Gil, C. u. a. (2014): Virtual Desktop Infrastructure (VDI) Technology : FI4VDI Project, in: Álvaro Rocha u. a. (Hrsg.) New Perspectives in Information Systems and Technologies, Volume 2, Bd. 276, Cham: Springer, S. 35–42

Rocha, Á. u. a. (Hrsg.) (2014): New Perspectives in Information Systems and Technologies, Volume 2, Bd. 276, Cham: Springer

Schawel, C. / Billing, F. (2004): Die Top 100 Management Tools : Das wichtigste Buch eines Managers, 1. Aufl, Wiesbaden: Gabler

Shen, G. / Huang, X. (Hrsg.) (2011): Advanced research on computer science and information engineering : International conference, CSIE 2011, Zhengzhou, China, May 21-22 2011, proceedings, 152-153, Heidelberg: Springer

Zehetmaier, J. (2011): Open Source Software in versicherungsfachliche Anwendungen, 1. Aufl, Karlsruhe: Verlag Versicherungswirtschaft GmbH

Verzeichnis der Internet- und Intranet-Quellen

Bittman, T. / Margevicius, M. / Dawson, P. (2014):
Magic Quadrant for x86 Server Virtualization Infrastructure, http://www.gartner.com/technology/reprints.do?id=1-1WR7CAC&ct=140703&st=sb, Abruf: 22.01.2015

Cupp Jr., D. A. (o. J.): Thinstation Linux Distro Maintainer / Developer, http://www.doncuppjr.net/index.php/projects, Abruf: 18.01.2015

Cupp Jr., D. A. (2014a): Getting Started with Thinstation, https://github.com/Thinstation/thinstation/wiki/Getting-Started-with-Thinstation, Abruf: 16.01.2015

Cupp Jr., D. A. (2014b): Thinstation : sourceforge, http://sourceforge.net/projects/thinstation/, Abruf: 19.01.2015

Cupp Jr., D. A. (2014c): Thinstation/thinstation - FAQ, https://github.com/Thinstation/thinstation/wiki/FAQ, Abruf: 20.01.2015

Gartner Inc. (2015): Gartner Magic Quadrant : Positioning Technology Players Within a Specific Market, http://www.gartner.com/technology/research/methodologies/research_mq.jsp, Abruf: 22.01.2015

GitHub Inc. (2015a): openthinclient-suite, https://github.com/openthinclient/openthinclient-suite/graphs/contributors, Abruf: 20.01.2015

GitHub Inc. (2015b): Thinstation : Contributions to 5.3-Stable, https://github.com/openthinclient/openthinclient-suite/graphs/contributors, Abruf: 20.01.2015

Henry, M. (1999): Intel Preboot Execution Environment, http://tools.ietf.org/id/draft-henry-remote-boot-protocol-00.txt, Abruf: 22.1.2015

IGEL Technology GmbH (2007): Anwenderbericht mit IGEL das Spiel gewinnen, 2007, https://www.igel.com/fileadmin/user/upload/documents/PDF_files/Case_Study_DE/CS_Borussia_DE.pdf, Abruf: 16.01.2015

IGEL Technology GmbH (2010): Wettbewerbsvorteil : Thin Clients im Versicherungswesen, https://www.igel.com/fileadmin/user/upload/documents/PDF_files/White_Paper_DE/WP_Insurance_99-DE-15-2.pdf, Abruf: 21.01.2015

IGEL Technology GmbH (2012): Thin Client Computing für das Versicherungswesen, https://www.igel.com/fileadmin/user/upload/documents/PDF_files/Case_Study_DE/IL_Insurance_155-DE-2-4.pdf, Abruf: 16.01.2015

IGEL Technology GmbH (2015): Was sind Thin Clients, https://www.igel.com/de/, Abruf: 16.01.2015

o. V. (2001): Virtual Private Networking: An overview, https://msdn.microsoft.com/en-us/library/bb742566.aspx, Abruf: 21.01.2015

o. V. (2014a): Thinstation, http://en.wikipedia.org/wiki/Thinstation, Abruf: 20.01.2015

o. V. (2014b): VDI und virtualisierte Anwendungen, http://www.pcwelt.de/ratgeber/VDI-und-virtualisierte-Anwendungen-1362295.html, Abruf: 17.01.2015

o. V. (2015): Logo Usage Guidelines, http://opensource.org/logo-usage-guidelines#The_OSI_Logo:_Usage_Guidelines, Abruf: 17.01.2015

Open Thin Client Alliance (2011): Founding members and sponsors, http://openthinclientalliance.org/index.php?option=com_content&view=article&id=2&Itemid=3, Abruf: 22.01.2015

openthinclient GmbH (o. J.): openthinclient® Anwendung Citrix ICA Client 13, http://www.openthinclient.net/epages/63830524.sf/de_DE/?ObjectPath=/Shops/63830524/Products/33864/, Abruf: 18.01.2015

openthinclient GmbH (2014a): Download openthinclient, http://openthinclient.org/de/download-openthinclient/, Abruf: 22.01.2015

openthinclient GmbH (2014b): features, http://features.openthinclient.org/, Abruf: 20.01.2015

openthinclient GmbH (2014c): Firmenprofil, http://openthinclient.com/wp-content/uploads/2012/12/openthinclient_Firmenprofil_de.pdf, Abruf: 19.01.2015

openthinclient GmbH (2014d): openthinclient Software Suite, http://openthinclient.com/wp-content/uploads/openthinclient_Software-Suite_de.pdf, Abruf: 20.01.2015

openthinclient GmbH (2014e): openthinclient WIKI : Einleitung, http://wiki.openthinclient.org/wiki/Einleitung, Abruf:19.01.2015

openthinclient GmbH (2014f): QuickStart, http://wiki.openthinclient.org/wiki/QuickStart, Abruf: 22.01.2015

openthinclient GmbH (2014g): Referenzen, http://openthinclient.com/referenzen/. Abruf: 16.01.2015

openthinclient GmbH (2014h): ThinClient Server Installation, http://wiki.openthinclient.org/wiki/ThinClient_Server_Installation, Abruf: 17.01.2015

openthinclient GmbH (2015): Funktionsprinzip, http://openthinclient.com/funktionsprinzip/, Abruf: 22.01.2015

Schnabel, P. (2014): VPN : Virtual Private Network, http://www.elektronik-kompendium.de/sites/net/0512041.htm, Abruf: 20.01.2015

VM Ware Inc. (2014): Performance Best Practices for VMware vSphere® 5.5, http://www.vmware.com/pdf/Perf_Best_Practices_vSphere5.5.pdf, Abruf: 22.01.2015

VM Ware Inc. (2015a): Build your own cloud infrastructure in your datacenter and remote site on VMware vSphere,

http://www.vmware.com/products/vsphere/features.html, Abruf: 22.01.2015

VM Ware Inc. (2015b): Products, http://www.vmware.com/products/, Abruf: 22.01.2015

VM Ware Inc. (2015c): VMware Player Pro, http://www.vmware.com/products/player/compare.html, Abruf: 22.01.2015

Wurm, Michaela (2013): Igel macht alte XP-Rechner zu Thin Clients, http://www.crn.de/server-clients/artikel-99424.html, Abruf: 16.01.2015

Gesprächsverzeichnis

Anonymisiert, A. (2015): Referat X , V, persönliches Gespräch am 09. Dezember 2015 in Stuttgart



Eine Marktanalyse über Open Source Dokumentationssysteme unter Berücksichtigung effizienten Wissensmanagements

Eine Empfehlung für die Anwendungsentwicklung der .Versicherung

Schriftliche Ausarbeitung
im Rahmen der Lehrveranstaltung „Integrationsseminar"
für das Kompetenzzentrum Open Source (KOS)

Vorgelegt von

Julia Flohr, Katharina Kowsky,
Franziska Lang, Magdalena Schwemmer,
Nina Strasser, Sabrina Tönhardt

am 04.02.2015

Fakultät Wirtschaft
Studiengang Wirtschaftsinformatik
WWI2012E

Inhaltsverzeichnis

Abkürzungsverzeichnis

ACL	Acess Control List
AWS	Amazon Web Services (Cloud Computing Service)
CAD	Computer-Aided Design
CMS	Content Management System
CRM	Customer Relationship Management
doc	document (Dateiformat)
DMS	Document Management System
Docx	document (Dateiformat)
gif	Graphics Interchange Format
GNU	GNU is not Unix
GPL	General Public License
HTML	HyperText Markup Language
IEEE	Institute of Electrical and Electronics Engineers
IP	Internetprotokoll
IT	Informationstechnologie
jpg	joint photographic group (Dateiformat)
jpeg	joint photographic experts group (Dateiformat)
KLM	Koninklijke Luchtvaart Maatschappij (königliche Luftfahrtgesellschaft)
K.O.	Knockout
KOS	Kompetenzzentrum Open Source
KVaG	Krankenversicherung auf Gegenseitigkeit
LKW	Lastkraftwagen
MB	Mega Byte
mp3	eigentlich MPEG Moving Picture Experts Group (Audio Dateiformat)
NASA	National Aeronautics and Space Administration
OS	Operation System
OSI	Open Source Initiative
PC	Personal Computer

PDF	Portable Document Format (Dateiformat)
PHP	Hypertext Preprocessor
PKW	Personenkraftwagen
png	portable network graphics (Dateiformat)
ppt	PowerPoint (Dateiformat)
pptx	PowerPoint (Dateiformat)
RAM	Random-Access Memory
RoK	Return on Knowledge
SAP	Systeme Anwendungen und Produkte in der Datenverarbeitung
SMW	Semantic Media Wiki
SOAP	Simple Object Access Protocol
SSL	Secure Sockets Layer
TCO	Total Cost of Ownership
txt	Text (Dateiformat)
UML	Unified Modelling Language
vs	versus
VW	Volkswagen
wav	WAVE (unkomprimiertes Audio Dateiformat)
wmv	windows media video (Dateiformat)
xls	Excel (Dateiformat)
XP	Experience
ZDF	Zweites Deutsches Fernsehen
zip	Zipper (Dateiformat)

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Das Kompetenzzentrum Open Source (KOS) hat in Kooperation mit der .Versicherung eine Aufgabenstellung über die Evaluation unterschiedlicher Open Source Dokumentationssysteme an die Duale Hochschule Baden-Württemberg Stuttgart vergeben. Im Rahmen dieser Projektaufgabe ist die zugrundeliegende Arbeit entstanden.

Ziel der vorliegenden Arbeit ist es, eine passende Dokumentationsplattform für die Abteilung der Anwendungsentwicklung der .Versicherung zu finden, die zu einem effektiveren Wissensmanagement beiträgt. Die Dokumentationsplattform soll in erster Linie die Bewahrung von Wissen gewährleisten. Die bisher genutzte Notesdatenbank, die schon einige Dokumente der Anwendungsentwicklungsabteilung beinhaltet, erfüllt jedoch nicht alle Kriterien, die benötigt werden, um effektives Wissensmanagement betreiben zu können. Ein grundsätzlich bestehendes Problem ist, dass die Notesdatenbank nicht von allen Mitarbeitern genutzt und anerkannt wird. Diese Arbeit beschäftigt sich folglich nicht ausschließlich mit der Findung eines neuen Tools, das allen Ansprüchen der .Versicherung entspricht sondern ebenso mit den Prozessen der Wissensbewahrung, -verteilung und -nutzung aus dem Wissensmanagement.

Im theoretischen Teil dieser Arbeit wird zunächst das Thema Wissen behandelt. Hier wird der Begriff Wissen erläutert und definiert. Für den späteren Verlauf ist es von großer Bedeutun zwischen implizitem und explizitem Wissen zu unterscheiden. Da sowohl der Begriff des Wissensmanagements, wie auch die einzelnen Prozessschritte und Bausteine des Wissensmanagements eine entscheidende Rolle für diese Arbeit spielen, wird im theoretischen Teil hier der Fokus gelegt.

Ein Exkurs in das Anforderungsmanagement zeigt wie wichtig die Formulierung der Anforderungen ist und unterstützt im späteren Verlauf der Arbeit das Vorgehen bei der Erstellung des Kriterienkatalogs.

Mehrere Experten- und Fachinterviews bilden die Grundlage des praktischen Teils. Um eine angemessene Empfehlung an die .Versicherung über Open Source Dokumentationssysteme abgeben zu können, muss zunächst eine IST-Analyse des aktuell genutzten Tools durchgeführt werden. Auf Basis der IST-Analyse wird eine SOLL-Analyse vorgenommen, die die Basis für die Erstellung eines Kriterienkatalogs bildet. In Rücksprache mit einem Vertreter der Anwendungsentwicklung der .Versicherung wird der Kriterienkatalog genehmigt und einzelne Kriterien sollen priorisiert werden. So kann im Anschluss eine umfangreiche Marktanalyse durchgeführt werden. Das Internet bietet hier die Grundlage, um Dokumentationstools ausfindig zu machen. Zu den wichtigsten Kriterien des neuen Tools zählen vor allem,

dass es Open Source Software ist, eine simple und schnell nachvollziehbare Navigation hat und eine äußerst gute Suchfunktion (nach Titeln, Autoren oder auch Stichworten) beinhaltet. So können schnell Tools gefunden werden, die in die engere Auswahl der Marktanalyse kommen. Hier wird davon ausgegangen, dass der Befund äußerst umfangreich sein wird. Aus diesem Grund ist der nächste Schritt der Marktanalyse Tools auszuwählen, die möglichst vielen Kriterien gerecht werden.

Als besonders wichtig ist die genaue Analyse der einzelnen Tools zu beurteilen. Hier werden fünf Tools näher betrachtet. Da in diesem Schritt davon ausgegangen wird, dass diese Tools bereits möglichst viele Kriterien erfüllen, geht es als nächstes darum, die Tools gegeneinander abzuwägen und Ausschlusskriterien zu finden, die die Auswahl der Tools einschränkt.

Ziel der Marktanalyse ist es, letztendlich eine Empfehlung für die .Versicherung für ein Dokumentationstool aussprechen zu können.

2 Theoretischer Hintergrund

In den folgenden Unterkapiteln werden für die Analyse relevante theoretische Hintergründe näher untersucht. Hierfür befassen sich die Autoren mit dem Thema Wissen und untersuchen dabei den Unterschied zwischen implizitem und explizitem Wissen, sowie die Theorien des Wissensmanagements. Ein Exkurs in die Theorien des Anforderungsmanagements zeigt dem Leser die Wichtigkeit eines qualitativ hochwertigen Kriterienkatalogs und zeigt Mittel auf, wie die Formulierung von Anforderungen optimal umgesetzt werden kann.

2.1 Wissen

Um im späteren Verlauf dieser Arbeit den Bezug zu Wissensmanagement herleiten zu können, muss an erster Stelle geklärt werden, was Wissen nach Definition ist und wieso es von solch hoher Bedeutung für Unternehmen ist. Grundsätzlich ist bereits an dieser Stelle festzuhalten, dass es keine allgemeingültige und fundierte Definition für Wissen gibt. Obwohl es im Alltag allgemeingebräuchlich ist und jeder annimmt, unter dem Wort Wissen direkt zu verstehen, was gemeint ist, so ist es doch äußerst schwierig, die genaue Bedeutung von Wissen allumfassend zu formulieren.
Schon in der Antike stellte Sokrates fest: „Umso mehr ich weiß, weiß ich, dass ich nichts weiß.“ Es wird deutlich, welch breiten Interpretationsspielraum die unterschiedlichen Definitionen hinterlassen.
Brockhaus definiert Wissen folgendermaßen: „1) alle Kenntnisse im Rahmen alltäglicher Handlungs- und Sachzusammenhänge (Alltagswissen); 2) im philosophischen Sinne die be-

gründete und begründbare (rationale) Erkenntnis im Unterschied zur Vermutung und Meinung oder zum Glauben. Wissen kann primär durch zufällige Beobachtung, durch systematische Erforschung (Experiment) oder deduzierende Erkenntnis gewonnen werden, sekundär durch lernende Aneignung von Wissensstoff."[1] Bei den kursierenden Definitionen von Wissen unterscheidet man in der Regel nach verschiedenen wissenschaftlichen Disziplinen.[2] So wird beispielsweise in der Philosophie, der Soziologie, der Psychologie und in der Informatik Wissen unterschiedlich definiert.

Um eine möglichst allgemeingültige Definition von Wissen entwickeln zu können, ist es wichtig, den Unterschied zwischen Zeichen, Daten, Information und Wissen zu verstehen. Hierzu dient Abbildung 1.

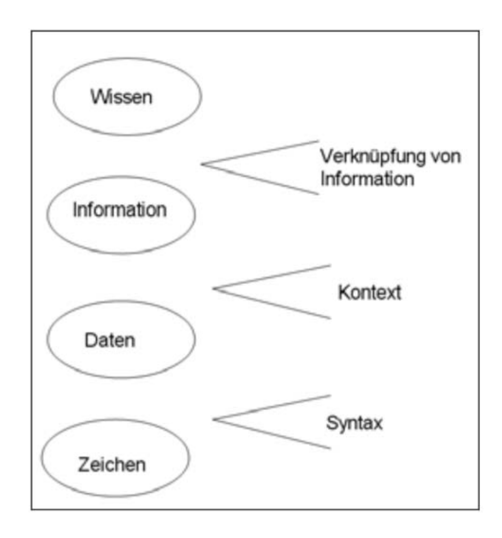


Abb. 1: Zeichen-Daten-Information-Wissen[3]

Abbildung 1 ist der Theorie der Informatik nachempfunden. So stellen in dieser Theorie auf dem Weg zum Wissen Zeichen die unterste Ebene der Unterscheidung dar.[4] Beispiele hierfür sind die Ziffer „5", das „ , "oder die Ziffer „7". Aus mehreren Zeichen, also beispielsweise einem Zeichenzusammenhang der Zeichen, wie in diesem Beispiel „5,7", ergeben Daten. Die Transformation von Daten zu Informationen ist je nach Wissenschaftsbereich anders ausgelegt. Hasler Roumois beschreibt die Transformation folgendermaßen: Immer, „wenn er [der Mensch, der Verfasser] ihre [die Daten] Relevanz für sein aktuelles (Wissens-)Bedürfnis erkennt, werden die Daten für diesen Menschen zu Informationen"[5]. Bringt ein Mensch also persönliches Interesse für Daten jeglicher Art auf, so werden diese Daten für ihn zu Informa-

[1] Der Brockhaus (1998)
[2] Vgl. Ahlert, M./Blaich, G./Spelsiek, J. (2006), S.36
[3] Entnommen aus: Gust von Loh, S. (2009)
[4] Vgl. Kenning, P./Schütte, R./Blaich, G. (2003), S. 32
[5] Hasler Roumois, U. (2007), S. 34

tionen. Wichtig hierfür ist, dass ein Individuum Daten in einen Kontext stellt.[6] Wissen wiederum entsteht nicht durch Informationen alleine. Das Individuum muss Informationen mit erlebten und eigenen Erfahrungen in Verbindung setzen, um so neues Wissen generieren zu können. In Verbindung setzen heißt, dass „die neuen Informationen mit dem schon vorhandenen Wissen gedanklich verknüpft werden."[7] Demnach entsteht Wissen lediglich im Kopf des Wissenden.

Insbesondere für Unternehmen entsteht an dieser Stelle die Problematik. Wissen ist eine extrem wertvolle Ressource von Unternehmen. Geschäftsprozesse und optimale Abläufe sind oftmals lediglich in den Köpfen der Mitarbeiter gespeichert. Verlässt ein Mitarbeiter aus unterschiedlichen Gründen das Unternehmen, so geht dieses Wissen verloren. Eben weil Wissen persönlich und erfahrungsgebunden ist, kann es nicht, ähnlich wie eine Datei auf dem Computer, weitergegeben oder abgespeichert werden.[8] Ebenso ist es zudem schwierig, erfahrungsgebundenes Wissen einfach zu erklären. Möchte beispielsweise Mitarbeiter X, der in naher Zukunft das Unternehmen verlässt, seinem Nachfolger, Mitarbeiter Y, erklären, wie er jeden Morgen den Kassenabschluss vom gestrigen Tag überprüft, so erweist sich dies als nicht so einfach wie angenommen. Den technischen Ablauf zu erläutern mag nicht sonderlich kompliziert sein, allerdings hat Mitarbeiter X persönlichen Ablauf, mit bestimmten Tastenkombinationen, Papierstapeln, die links statt rechts von ihm liegen, zu erklären und zu begründen, warum dies so ist. Die Erläuterung der persönlichen Arbeitsweise ergibt sich als äußerst schwierig. Mitarbeiter X hat also für den gewöhnlichen Prozess den schnellsten Ablauf gefunden, weil er persönliche Erfahrungen mit Informationen in Verbindung gesetzt hat.

2.1.1 Implizites vs. explizites Wissen

An dieser Stelle ist der Unterschied zwischen impliziten und explizitem Wissen zu erläutern.

Wissen, das anderen nicht erklärt werden kann, so wie im Falle von Mitarbeiter X, wird grundsätzlich implizites Wissen genannt. Implizites Wissen basiert auf Erfahrungen. Ein Mechanikermeister kann aufgrund seiner Erfahrung viel schneller den Fehler am Motor eines Kundenfahrzeugs feststellen als ein Lehrling es könnte.[9] In diesem Sinne erweist es sich als äußerst schwierig, implizites Wissen in Worte zu fassen. Laut Polanyi existiert implizites Wissen als Hintergrundwissen.[10]

[6] Vgl. Gust von Loh, S. (2009), S.12
[7] Reiber, W. (2013), S.38
[8] Vgl. Reiber, W. (2013), S.38
[9] Vgl. Reiber, W. (2013), S.39
[10] Vgl. Gust von Loh, S. (2009), S.15

Laut Schwaninger lautet die Definition für implizites Wissen wie folgt: „Implizites Wissen besteht in der Fähigkeit, Distinktionen und Selektionen weitgehend intuitiv, laufend zu treffen sowie diese in praktische Handlungen umzusetzen."[11]

Implizites Wissen	**Explizites Wissen**
Erfahrungswissen (Körper)	Verstandeswissen (Geist)
Gleichzeitiges Wissen (hier und jetzt)	Sequentielles Wissen (da und damals)
Analoges Wissen (Praxis)	Digitales Wissen (Theorie)
Unartikulierbar	Verbalisierbarkeit
Speicherung in Datenbanken und anderen Medien nicht möglich	Speicherung in Datenbanken und anderen Medien
Weitergabe durch Zusammenarbeit und Beobachtung	Weitergabe durch Datenbanken und andere Medien
Erfahrungswissen/unbewusstes Wissen	Leicht zu strukturieren

Tabelle 1: Implizites vs. explizites Wissen[12][13]

Wie man anhand von Tabelle 1 bereits erkennen kann, ist explizites Wissen um einiges leichter zu erläutern. Explizites Wissen ist Wissen, das sich der Mensch selbst und bewusst angeeignet hat. Zudem kann explizites Wissen abgespeichert werden. Schwaninger definiert explizites Wissen folgendermaßen: „Explizites Wissen umfasst Inhalte, die aufgrund bereits getroffener Unterscheidungen (Distinktionen) und durch Auswahlvorgänge (Selektionen) zustande gekommen sind."[14] Explizites Wissen kann beispielsweise das Fachwissen in einem Buch sein.

Demzufolge ist explizites Wissen um einiges leichter weiterzugeben als implizites Wissen. Implizites Wissen ist äußerst schwer durch den Wissenden alleine an eine andere Person weiterzugeben. Die lernende Person muss durch Beobachtung des impliziten Wissens ihre eigenen Erfahrungen machen und so für sich persönlich feststellen, wie der Wissende sein implizites Wissen aufgebaut hat. Eine weitere Schwierigkeit bei der Weitergabe von implizitem Wissen ist die Aufnahme des Lernenden von Gesagtem. Dies bedeutet, dass der Vermittelnde beim Versuch sein implizites Wissen weiterzugeben, nicht die Kontrolle darüber

[11] Schwaninger, M. (2000), S.4
[12] Mit Änderungen entnommen aus: Frey-Luxemburger, M. (2014), S.19
[13] Mit Änderungen entnommen aus: Gust von Loh, S. (2009), S.16
[14] Schwaninger, M. (2000), S.3

hat, was der Lernende hört oder versteht.[15] Ist die Erklärung des Vermittelten also schlüssig und logisch, so muss dies nicht dasselbe für den Lernenden bedeuten.

Zum weiteren Verständnis von Wissen ist grundsätzlich festzuhalten, dass Wissen nach Polanyi immer aus implizitem und explizitem Wissen besteht. Nachdem nun die Grundbegriffe von Wissen und die Unterscheidung zwischen implizitem und explizitem Wissen ausgeführt wurden, kann eine für diese Arbeit endgültige Definition von Wissen aufgestellt werden.

„Wissen bezeichnet die Gesamtheit der Kenntnisse und Fähigkeiten, die Individuen zur Lösung von Problemen einsetzen. Dies umfasst sowohl theoretische Erkenntnisse als auch praktische Alltagsregeln und Handlungsanweisungen. Wissen stützt sich auf Daten und Informationen, ist im Gegensatz zu diesen jedoch immer an Personen gebunden. Es wird von Individuen konstruiert und repräsentiert deren Erwartungen über Ursachen-Wirkungs-Zusammenhänge."[16]

Es ist unumstritten, dass Wissen im heutigen Zeitalter die mit Abstand wichtigste Ressource eines jeden Unternehmens ist. So wird behauptet, dass bereits heute rund dreiviertel des generierten Mehrwertes eines Unternehmens auf spezifisches Wissen zurückzuführen ist.[17] Unternehmen leben von impliziten Wissen und müssen dieses nach Möglichkeit – wie selbstverständlich auch explizites Wissen – festhalten. Wissen, das sich in der Praxis bewährt hat, gilt als äußerst kostbar. Geht es einmal verloren, so ist es für immer fort. Durchaus kann neues Wissen generiert werden, dennoch gehen kostbare Ressourcen verloren, die nicht wieder aufbringbar sind. Zudem lebt Wissen von Wissen. Somit ist „Wissen die einzige Ressource, welche sich durch Gebrauch vermehrt."[18] Durch Wissen können in Unternehmen die Unternehmensziele verfolgt und erreicht werden. Wissen ist die Grundlage zur Problemlösung, wie z.B. von Kundenproblemen. Somit wird der Wert von Wissen auch danach beurteilt, inwieweit es sich in der Praxis bewährt.[19]

2.1.2 Wissensmanagement

Wissensmanagement spielt im Zeitalter der Globalisierung eine enorm wichtige Rolle für Unternehmen. Oftmals wird der Begriff Wissensmanagement automatisch mit reiner technologischer Systemunterstützung in Verbindung gesetzt. Dies ist allerdings falsch. Grundsätzlich hat das reine Wissensmanagement erst einmal die Aufgabe, Unternehmen dabei zu unter-

[15] Reiber, W. (2013), S.34
[16] Probst, G. J. B./Raub, S./Romhardt, K. (2006), S.23
[17] Vgl. Probst, G./Raub, S./Romhardt, K. (2012), S.3
[18] Probst, G./Raub, S./Romhardt, K. (2012), S.2
[19] Vgl. Reiber, W. (2013), S.37

stützen, Informationen und Wissen strategisch und effektiv einzusetzen.[20] Es bedeutet zudem, äußerst vorsichtig und bewusst mit Wissen umzugehen, um dieses zu schützen – es soll nicht an die falschen Unternehmen ausgetragen werden -, es soll aber auch innerhalb des Unternehmens verbreitet und vor allem erweitert werden. Broßmann und Mödinger treffen diesbezüglich eine passende Aussage: „[..] der spezielle Umgang mit dem besonderen Gut ‚Wissen' einer ‚empfindlichen exotischen Pflanze' gleicht, die es einerseits zu schützen und zu hegen gilt, damit sie die volle Wirkung entfalten kann, die sich andererseits aber auch den härteren klimatischen Bedingungen und Herausforderungen stellen muss, wie es das Wirtschaften mit Gütern in Unternehmen und Organisationen nach sich zieht."[21] Es gilt demnach nicht nur, Wissen zu organisieren und damit zu schützen, sondern es ebenso auszubauen, um dem Unternehmen so einen Wettbewerbsvorteil schaffen zu können. Da Wissensmanagement darauf abzielt, die Unternehmensziele zu verfolgen, ist es auf gleicher Ebene mit der Unternehmensstrategie zu setzen. Beide hängen in unterschiedlichen Punkten stark zusammen und sind ebenso abhängig voneinander. Nicht nur Wissensmanagement/Wissensstrategien folgen der Unternehmensstrategie, es/sie wird/werden auch von ihr abgeleitet. Umgekehrt gilt für die Unternehmensstrategie dasselbe. Das Wissensmanagement kann sich in seiner Strategie aber auch auf unterschiedliche Funktionen beschränken oder spezialisieren. So kann, wie bereits erwähnt, Wissensmanagement mit der Unternehmensstrategie gleichgesetzt werden. Es kann sich allerdings auch auf das Management des intellektuellen Kapitals beschränken, es kann kundenorientiert sein, zum Best-Practice-Sharing dienen, als Wissensgenerierung und Innovation wirken oder andere Aufgaben übernehmen.[22] Welche Strategie des Wissensmanagements ein Unternehmen wählt, hängt von seinen Zielen ab. Diese Arbeit wird sich auf das Wissensmanagement als Unternehmensstrategie beschränken.

Nachdem die grundlegenden Aufgaben und die Ausrichtungen des Wissensmanagements genannt wurden, wird auf die Elemente und Grundprozesse eingegangen. Mindestens drei Elemente müssen vorhanden sein, um Wissensmanagement zu beschreiben. Das TOM-Modell in Abbildung 2 verdeutlicht diese.

[20] Vgl. Gretsch, S. M. (2014), S.26
[21] Broßmann, M./Mödinger, W. (2008), S.18
[22] Vgl. Frey-Luxemburger, M. (2014), S.33

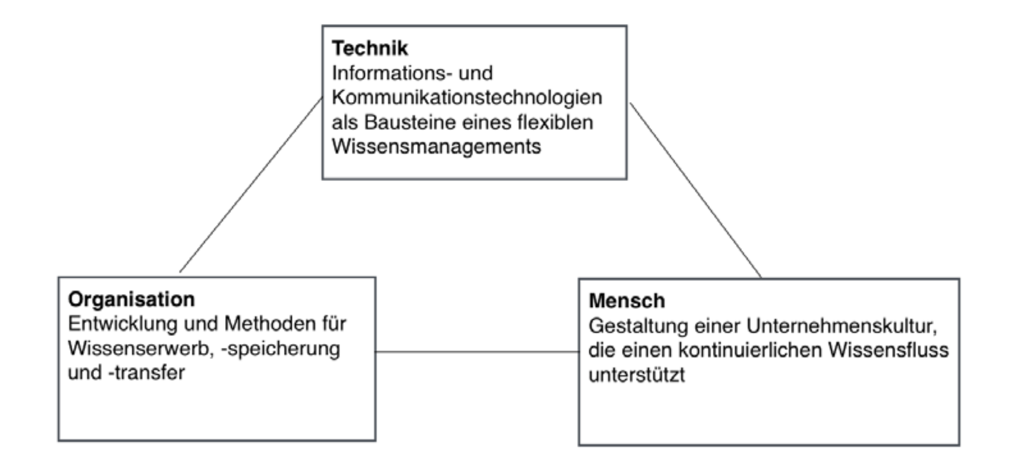


Abb. 2: TOM-Modell[23]

Das TOM-Modell bezieht die drei wichtigsten Faktoren für das Wissensmanagement mit ein. Technik, die Organisation und der Mensch machen Wissensmanagement erst möglich. Diese Faktoren werden im Folgenden etwas näher betrachtet.

Das Thema Technik beschäftigt sich in diesem Sinne mit der Speicherung und der Ablage von (speziell) explizitem Wissen. An dieser Stelle muss zunächst der Unterschied zwischen Informationsmanagement und Wissensmanagement erläutert werden.
Informationsmanagement zeichnet sich im Gegensatz zu Wissensmanagement tatsächlich dadurch aus, dass es durch Technologie unterstützt wird bzw. dass die Hauptaufgabe des Informationsmanagements darin besteht, Informations- und Kommunikationstechnologien zur Verfügung zu stellen.[24] Rückblickend auf den am Anfang der Arbeit beschriebenen Unterschied zwischen Information und Wissen, so wird hier deutlich, dass das Informationsmanagement lediglich – wie das Wort an sich schon sagt – darauf abzielt, Informationen, die noch kein Wissen sind, zu sammeln und beispielsweise auf einer Datenbank abzulegen. Trotzdem benötigt ein Unternehmen zum Betreiben von Wissensmanagement ebenso eine Datenbank oder Ähnliches, auf der insbesondere explizites, also niederschreibbares Wissen abgelegt und gepflegt werden kann.

Als zweiter entscheidender Faktor zählt die Organisation, also das Unternehmen. Es ist die Aufgabe der Organisation, Wissensmanagement in das Alltagsgeschäft des Unternehmens zu etablieren. Dies bedeutet, dass Wissensmanagement bei seiner offiziellen Einführung – viele Unternehmen betreiben bereits Wissensmanagement, ohne es bewusst wahrzunehmen

[23] Mit Änderungen entnommen aus: Lucko, S./Trauner, B. (2005), S.24
[24] Vgl. Gust von Loh, S. (2009), S.25

– von den Mitarbeitern akzeptiert und betrieben werden muss. Diese Kommunikation und Vermittlung ist die Aufgabe der Organisation, bzw. des Managements.[25]

Die Grenzen zwischen Organisation und Mensch sind fließend. Auch wenn die Organisation versucht, Wissensmanagement attraktiv für die Mitarbeiter zu machen, so ist es letztendlich der Mensch bzw. Der Mitarbeiter, der darüber entscheidet, ob er bereit ist, sein Wissen zu teilen und zu verbreiten oder nicht. Der Mensch bildet nicht nur im TOM-Modell sondern im gesamten Wissensmanagement den Mittelpunkt, da die Generierung und Evaluierung neuen Wissens nur durch ihn erfolgen kann.[26]

Es ist wichtig nie außer Betracht zu lassen, dass Wissensmanagement stark durch die Einschränkungen, die ein jedes Individuum mit sich bringt, behindert werden kann. Der Mensch/Mitarbeiter ist der interpretierende Handlungsträger, der aus Informationen Wissen schafft. Ohne seine geistige Leistung entstünde für das Unternehmen kein Wissen. Aus diesem Grund sehen Seiler und Reinmann Wissensmanagement als „eine[r] integrative[n] Aufgabe, zwei fundamentale Arten des Managements von Wissen zusammenbringen muss: das Management von objektivierten (öffentlichem) Wissen im klassischen Sinne der Planung, Steuerung und Kontrolle sowie das Management von idiosynkratischem (personalem) Wissen in Sinn der Förderung menschlicher Fähigkeiten, Bereitschaften, Austausch- und Gestaltungsprozessen.“[27]

Demzufolge ist an dieser Stelle festzuhalten, dass die Ziele von Wissensmanagement die Steigerung der Innovationsfähigkeit, das organisationale Lernen und eine bessere Nutzung des vorhandenen Wissens innerhalb einer Organisation sind. Das Oberziel hierzu ist, die Wettbewerbsfähigkeit eines Unternehmens zu steigern und damit einen Beitrag zum langfristigen Erfolg zu liefern.[28] Entscheidende Faktoren zur Lieferung dieser Ziele sind Mensch, Organisation und Technik.

Die **Kernprozesse des Wissensmanagements** haben in einer praxisnahen Aktionsforschung Probst, Raub und Romhardt identifiziert. Das von ihnen erstellte Modell mit den sechs Kernprozessen hat große Popularität im deutschsprachigen Europa erlangen können und gilt seither als Quasi-Standard.[29] Die sechs Grundprozesse oder auch Bausteine des Wissensmanagements beschreiben Wissensidentifikation, Wissenserwerb, Wissensentwicklung, Wissens(ver)teilung, Wissensnutzung und Wissensbewahrung. Dieser operative Kreislauf wird ergänzt durch den Management-Kreislauf mit den Bausteinen Wissensziele und

[25] Vgl. Gust von Loh, S. (2009), S.29
[26] Vgl. Kusterer, S. (2008), S.31
[27] Reinmann, G./Mandl, H. (2004), S.11ff.
[28] Vgl. Ahlert, M./Blaich, G./Spelsiek, J. (2006), S.55
[29] Vgl. Gehle, M. (2006), S.47

Wissensbewertung. Der gesamte Kreislauf wird als der Prozess des Wissensmanagements bezeichnet. Die einzelnen Bausteine sind abhängig voneinander und sollen somit nicht getrennt von einer betrachtet werden.[30] Diese sechs Bausteine erwiesen sich nach der engen Zusammenarbeit von Probst, Raub und Romhardt mit verschiedenen Unternehmen als Themen mit der größten praktischen Relevanz in Zusammenhang mit Wissen. So wurden praktische Probleme identifiziert, die einen Bezug zum Thema Wissen aufzuweisen hatten, aus denen dann die sechs Bausteine entstanden sind.[31] In Abbildung 3 verdeutlicht der äußere Kasten den Management-Kreislauf, während der innere Kasten den operativen Kreislauf verdeutlicht. Ein Durchlauf des Prozesses ist bei dem Stichwort Feedback im Management-Kreis zwischen den Bausteinen Wissensziele und –bewertung geschehen.

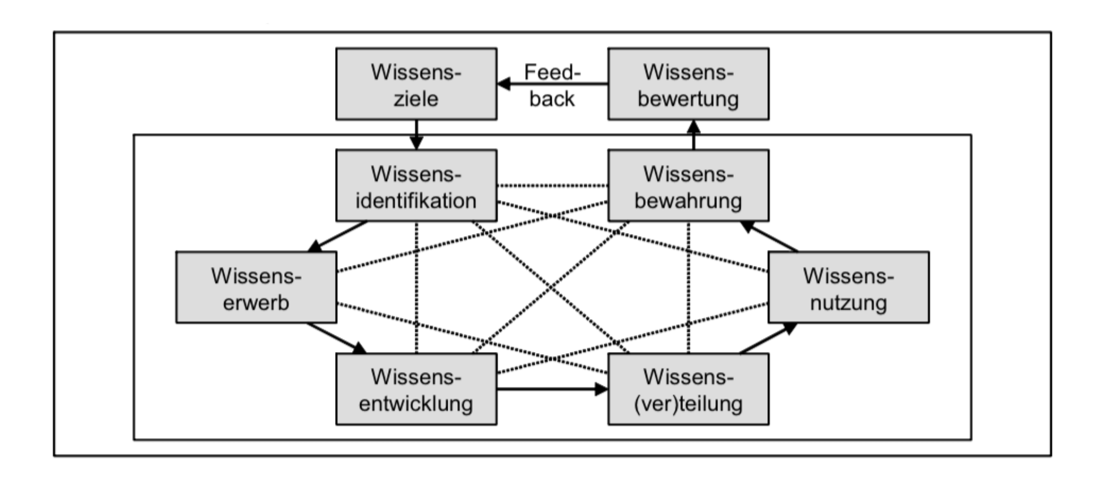


Abb. 3: Bausteine des Wissensmanagements[32]

Wissensziele:

„Wissensziele geben den Aktivitäten des Wissensmanagements eine Richtung."[33] Bei den Wissenszielen wird unter normativen, strategischen und operativen Zielen unterschieden. Die normativen Ziele beschäftigen sich mit der grundsätzlichen Auseinandersetzung mit Wissen. Sie bilden den unternehmenspolitischen Rahmen und die Grundlage, um überhaupt mit Wissensmanagement arbeiten zu können.[34] Strategische Ziele legen fest, welche Kompetenzen an Wissen für die Zukunft benötigt werden, um Unternehmensziele zu erreichen.[35]

[30] Vgl. Gretsch, S. M. (2014), S.28
[31] Probst, G. J. B./Raub, S./Romhardt, K. (2006), S.27f.
[32] Mit Änderungen entnommen aus: Probst, G. J. B./Raub, S./Romhardt, K. (2006), S.32
[33] Probst, G. J. B./Raub, S./Romhardt, K. (2006), S. 31
[34] Vgl. Kusterer, S. (2008), S.33
[35] Vgl. Gehle, M. (2006), S.49

Sie legen somit die angestrebten Ziele für die Zukunft fest. Operative Ziele dienen der Unterstützung der normativen und strategischen Ziele.[36] Sie dienen also zur Umsetzung.

Wissensbewertung:

Dieser Baustein ist der wohl wichtigste des gesamten Prozesses und zugleich auch der am schwierigsten umsetzbare. Wie im Controlling Kennzahlen generiert werden, um gewisse Qualitäten zu beweisen, so ist dies auch für die Wissensbewertung vorgesehen. Jedoch gibt es heutzutage noch keine bewährten Methoden, die es fundiert zulassen, das vorhandene Wissen zu bewerten. Insbesondere liefert sich hier weiter eine besondere Schwierigkeit in Bezug auf implizites Wissen. Für Wissen und Fähigkeiten existieren bis heute nur wenig brauchbare Indikatoren und Messverfahren.[37] Trotzdem ist es von enormer Bedeutung, das vorhandene Wissen zu messen, „denn erst wenn die Produktivität des Wissens in der Art eines RoK (Return on Knowledge) gemessen werden kann, erhält es die angemessene Aufmerksamkeit des Managements“.[38]

Wissensidentifikation:

Die Wissensidentifikation hat zum Ziel, das bereits vorhandene Wissen im Unternehmen zu analysieren. Es gilt, einen Überblick über interne und externe Daten, Informationen und Fähigkeiten zu behalten. Im Vordergrund steht hier die Transparenz für Mitarbeiter und Management. Ohne diese Transparenz können Ineffizienzen oder Doppelarbeiten auftreten, welche unnötig Zeit und Aufwand kosten.[39] Zu beachten bei der Wissensidentifikation ist nicht nur das explizite Wissen sondern ebenso das implizite Wissen. Hier gilt es Methoden anzuwenden, um auf implizites Wissen hinzuweisen. Hierzu dienen beispielsweise Expertenverzeichnisse, Wissenslandkarten oder Best-Practices.[40]

Wissenserwerb:

Neues, noch nicht im Unternehmen vorhandenes Wissen wird beim Wissenserwerb extern erworben oder akquiriert. Hierbei kann auf Beziehungen gesetzt werden, wie zum Beispiel zum Kunden oder zu Lieferanten oder Unternehmen in Kooperation. Es können zudem aber auch neue Mitarbeiter oder Experten eingestellt werden oder besonders innovative Unternehmen werden akquiriert.[41] Ein Grund für den Wissenserwerb durch externe Quellen ist,

[36] Probst, G. J. B./Raub, S./Romhardt, K. (2006), S.31
[37] Gehle, M. (2006), S. 49
[38] Gehle, M. (2006), S.49
[39] Vgl. Kusterer, S. (2008), S.33
[40] Vgl. Gehle, M. (2006), S.50
[41] Vgl. Probst, G. J. B./Raub, S./Romhardt, K.. (2006), S.29

dass es „nicht immer möglich oder effizient [ist], bestimmtes Wissen selbst zu entwickeln“.[42] Der Vorteil von Wissenserwerb ist, dass schnell mangelndes Expertenwissen herangezogen werden kann. Eine Problematik könnte jedoch sein, dass es zu einer „Abwehrreaktion auf Grund der Fremdartigkeit des Wissens“[43] kommt.

Wissensentwicklung:

Die Wissensentwicklung ist der komplementäre Baustein zum Wissenserwerb.[44] Bezogen wird sich in der Wissensentwicklung – wie vielleicht angenommen – nicht nur auf die Entwicklung und Forschung in ‚herkömmlichen‘ Gebieten wie den Naturwissenschaften. Die Wissensentwicklung strebt vor allem an, in allen Bereichen des Unternehmens Innovationen und Neuartigkeiten zu fördern. Hierzu zählen sämtliche organisational Lernprozesse, wie auch soziale Phänomene.[45] Hauptaufgabe der Wissensentwicklung ist es, mit der Kreativität der Mitarbeiter umzugehen. Dazu zählen die Sichtweite nicht einzuschränken, den Mitarbeitern mit Offenheit, Vertrauen und einem angemessenen Arbeitsklima entgegen zu kommen und so die Kreativität zu fördern. Primäres Ziel der Wissensentwicklung ist „die Kreierung intern und extern noch nicht existierender Fähigkeiten“[46]. Für die optimale Wissensentwicklung ist im Idealfall eine wenig hierarchische Organisation mit einer offenen Unternehmenskultur von Vorteil. Dadurch entsteht ein einfacherer Informationsaustausch zwischen den Mitarbeitern und Organisationseinheiten.[47]

Wissens(ver)teilung:

Ohne die Wissensverteilung würde Wissensmanagement nicht funktionieren. Das vorhandene und identifizierte Wissen muss an die richtigen Mitarbeiter und Organisationen weitergegeben werden. Schwierig hierbei ist, die richtigen Empfänger festzustellen. Es muss einerseits sichergestellt werden, dass eine gewisse Transparenz herrscht, sodass jeder Mitarbeiter Zugriff auf den Wissensbestand des Unternehmens hat. Andererseits verschafft zu viel ‚Material‘ eine eventuelle Überflutung und damit Überforderung an Wissen. In diesem Fall hat der Einsatz von Wissensmanagement einen kontraproduktiven Effekt.[48] Wissen erfährt an dieser Stelle eine Art ‚Entwertung‘, weshalb sich von Anfang an die Frage zu stellen ist,

[42] Gehle, M. (2006), S.51
[43] Gehle, M. (2006), S.51
[44] Kusterer, S. (2008), S.35
[45] Gehle, M. (2006), S.51
[46] Probst, G. J. B./Raub, S./Romhardt, K. (2006), S.29
[47] Vgl. Gust von Loh, S. (2009), S.22
[48] Vgl. Kusterer, S. (2008), S.35

wer was in welchem Umfang wissen soll oder kann und wie man die Prozesse der Wissensverteilung erleichtert.[49]

Wissensnutzung:

Die Wissensnutzung ist der wichtigste Baustein des Modells. Hierbei geht es darum, dass die Mitarbeiter das zur Verfügung gestellte Wissen auch nutzen. Denn die Produktion, Sortierung, Speicherung etc. reicht noch nicht aus, um daraus Wettbewerbsvorteile zu generieren.[50] Hierzu zählt wieder die Motivation der Mitarbeiter. Nicht zu vergessen ist im gesamten Prozess, dass der Mensch immer der Mittelpunkt ist. Aus diesem Grund ist ebenso darauf zu achten, dass Wissen eine gewisse Qualität aufweist, die es für die Mitarbeiter attraktiv macht. Problematisch kann hierbei sein, die Qualität des Wissens zu messen. Hinzu kommt, dass unterschiedliche Mitarbeiter die Qualität unterschiedlich bewerten.[51]

Wissensbewahrung:

Der letzte Punkt des operativen Kreislaufes ist die Wissensbewahrung. Insbesondere explizites Wissen kann in Dokumentenmanagementsystemen bewahrt werden. Hierzu zählen Datenbanken, Wikis oder ähnliche Systeme. Bei solchen Systemen steht vor allem der Schutz des Wissens im Vordergrund. Die Wissensbewahrung lässt sich in einem dreistufigen Prozess abbilden.



Abb. 4: Die Hauptprozesse der Wissensbewahrung[52]

Das Selegieren (Selektion) dient der Auswahl des bewahrungswürdigen Wissens. Dieses Wissen muss anschließend gespeichert werden, damit es Unternehmen dauerhaft zur Verfügung steht. Dies kann auf elektronischen Datenträgern, Datenbanken oder auch schriftlich geschehen. Der letzte Schritt ist die Aktualisierung. Wissen kann ergänzt oder gar erneuert werden. Damit das Wissen immer aktuell und auf dem neuesten Stand ist, muss eine kontinuierliche Aktualisierung des Wissens stattfinden.[53] Schwierig hierbei ist die Bewertung des ‚richtigen' Wissens.

[49] Gehle, M. (2006), S.54f.
[50] Gehle, M. (2006), S.56
[51] Vgl. Kusterer, S. (2008), S.36
[52] Mit Änderungen entnommen aus: Probst, G. J. B./Raub, S./Romhardt, K. (2006), S.193
[53] Vgl. Kusterer, S. (2008), S.36

Die Wissensbewahrung lebt nicht nur von der Pflege und Aktualisierung des Wissens, es muss zudem gewährleistet sein, dass jeder Mitarbeiter bereit ist, sein Wissen zu teilen. Eine einprägsame Aussage hat Blair diesbezüglich getroffen: „Every afternoon our corporate knowledge walks out the door and I hope to God they'll be back tomorrow“[54].

In vielen Werken der Literatur wird behauptet, dass Wissens nicht gemanagt werden kann.[55] So wäre der Baustein der Wissensbewahrung eher als Informationsbewahrung und somit als Informationsmanagement zu sehen. Jedoch ist Informationsmanagement als Teil des Wissensmanagements zu betrachten. Denn Technik ist ein Element des Wissensmanagements (siehe Abbildung 2). Demzufolge müssen Informations- und Wissensmanagement im dauerhaften Austausch stehen, um explizites Wissen abrufbar zu machen, aber auch, um implizites Wissen zu entwickeln.[56]

2.2 Anforderungsmanagement

Die .Versicherung hat gewisse Vorstellungen und Ansprüche bezüglich des firmeninternen Wissensmanagements.

Ziel dieses Projektes ist es, eine Empfehlung über die am Markt verfügbare Open Source Software, die sich als Dokumentationssystem zur Wissensbewahrung eignet, auszusprechen. Da es eine Vielzahl von Open Source Software in diesem Bereich gibt, muss expliziter definiert werden, was die Anforderungen des Unternehmens und des KOS an eine solche Software sind.

Ein Exkurs in das Anforderungsmanagement sorgt für ein strukturiertes Vorgehen in der weiteren Arbeit. Anforderungsmanagement ist Teil des Projektmanagements und wird oft in Bezug auf die Notwendigkeit unterschätzt.

An dieser Stelle ist hervorzuheben, dass zwischen Projektleiter und Anforderungsmanager unterschieden werden muss. Es handelt sich um zwei verschiedene Managementdisziplinen, welche sich jeweils ergänzen. Innerhalb des Anforderungsmanagements gibt es einen Anforderungsmanager, welcher in Kontakt zu allen Projektbeteiligten steht.

Im ersten Schritt hat der Anforderungsmanager die Aufgabe, herauszufinden welche Personen an dem Projekt beteiligt sind, um dann die Inhaber der Anforderungen zu definieren. Anschließend wird durch den Anforderungsmanager eine Übersicht erstellt in der sowohl bewusste als auch unbewusste Erwartungen und Vorstellungen über die Lösung zusam-

[54] Blair, D. C. (2002), S. 1021
[55] Vgl. Mujan, D. (2006), S. 58ff.
[56] Al-Hawamdeh, S. (2003), S.22

mengetragen werden. Diese wird dann so aufbereitet, dass daraus ein Gesamtbild für alle Beteiligten hervor geht.
Werden die Erwartungen zu spät kommuniziert oder werden Widersprüche zu spät aufgedeckt, so kann das erhebliche Schäden für den Erfolg des Projektes und für den Erfolg der Lösung haben.
Kommunikation ist ein sehr wichtiges Stichwort innerhalb des Anforderungsmanagements. Oft ist der Erfolg beeinflusst durch Sprachbarrieren, die zwischen unterschiedlichen Abteilungen liegen. Je spezifischer die Anforderung formuliert werden soll, desto mehr fachspezifische Begriffe werden verwendet. Jedoch ist beispielsweise die fachfremde Sprache des IT-Spezialisten nicht für alle verständlich und es muss eine einheitlich verständliche Sprache identifiziert werden. Daher gilt es Wege zu finden, mit denen Anforderungen verfasst werden können. Hier greift das Anforderungsmanagement mit dem Einsatz von Methoden, die die Qualität von Lastenheften sicherstellt.
Um das Anforderungsmanagement strukturiert anzugehen, ist es sinnvoll Arbeitsschritte zu definieren, die notwendig sind für die Erkennung und Verwaltung von Anforderungen. Es ist zu klären, welche Personen in welchen Arbeitsschritten welche Rolle haben. Diesen Schritt bezeichnet man als Erstellung eines **Vorgehensmodells**. Aus finanziellen Gründen wird dieser Schritt vor allem bei mittelständischen Unternehmen vernachlässigt und findet sich in abgewandelter Form in der Konzepterstellung wieder.
Werden in dieser Phase nicht alle erforderlichen Aspekte betrachtet, so können Lücken entstehen, die sich als Gefahr für das Projekt herausstellen können, da sie Raum für Interpretation lassen. Dies kann zu unterschiedlichen Erwartungshaltungen an die Lösung führen und bei Projektabschluss Unzufriedenheit zur Folge haben.
Für einen hohen Effekt des Anforderungsmanagements ist es wichtig, von Seiten des Managements aus unterschiedlichen Managementebenen unterstützt zu werden. Neben der Verfügbarkeit von erfahrenen Spezialisten im Bereich Anforderungsmanagement ist es erforderlich, Zeit und Geld zur Verfügung zu haben, um ein Vorgehensmodell zu erstellen.
Um die Problematik der Sprachbarrieren aufzugreifen, zeigt das Zitat von Thomas Niebisch sehr gut die Problematik auf: „Wenn im Eierkuchen Eier verarbeitet werden, was steckt dann im Sand-, Marmor oder gar Hundekuchen? Ist man masochistisch veranlagt, wenn man Verlangen nach einem Bienenstich verspürt? Zählt das Verspeisen von Hamburgern und Berlinern bereits zum Kannibalismus?“[57] Zunächst scheinen die Beispiele sehr weit hergeholt, jedoch zeigen sie plakativ die Gefahr von Missverständnissen auf Grund leichtfertiger Kommunikation. Im Alltag haben wir kaum Schwierigkeiten, da sich aus dem Kontext ableiten lässt, ob beispielsweise bei einem Berliner von einem Bewohner der Stadt Berlin die Rede

[57] Niebisch (2013), S. 13

ist, oder von einem Gebäck. Fehlt Kommunikation jedoch innerhalb eines IT Projekts, so kann es zu Irrtümern und falschen Interpretationen kommen.
Damit diese Gefahr umgangen wird, sieht das Anforderungsmanagement eine Überprüfung aller getätigten Aussagen vor. Die innerhalb des Projektes getroffenen Aussagen werden daher auf mutmaßliche Auslegungen und Interpretationen geprüft. Lässt eine Aussage unterschiedliche Deutungen zu, so wird die Aussage von verschiedenen Hören auch unterschiedlich ausgelegt.
Die Notwendigkeit, Anforderungen unmissverständlich zu formulieren wird immer deutlicher, denn „Sagen ist nicht Meinen und Hören ist nicht gleich Verstehen."[58] [59]

Der erste Schritt des Vorgehensmodells ist das Definieren von Projektbeteiligten, den sogenannten Stakeholdern. Zu ihnen soll ein Kontakt aufgebaut werden und die IST-Situation soll dann herausgearbeitet werden. Für die gesammelten Informationen gilt es dann, diese in eine Struktur zu bringen, die von allen Projektbeteiligten verstanden wird und einem Standard entspricht. Hier bietet es sich an, mit der Unified Modeling Language (UML) ein Use-Case-Diagramm zu erstellen.
UML hilft, die Realität in vereinfachter Form abzubilden und bietet die Möglichkeit, verschiedene Modelle aus simplen Grundbausteinen zu erstellen. Dazu gehören beispielsweise Klassen, Kollaborationen, Interfaces, Abhängigkeiten, Knoten, Komponenten, sowie Assoziationen und Generalisierungen.[60] Use-Case-Diagramme werden sehr häufig im Bereich der Anforderungsanalyse eingesetzt, denn sie bieten eine verständliche Basis für die Kommunikation unter Entwicklern, Anwendern und Analytikern.[61]
Wenn die IST-Situation für alle Beteiligten verständlich aufbereitet ist, kann zum nächsten Schritt übergegangen werden: Das Definieren von Arbeitspaketen. Durch das Definieren von Arbeitspaketen soll klar werden, welche Tätigkeiten notwendig sind für das Ermitteln und welche Anforderungen an die Lösung gestellt sind. Die verschiedenen Tätigkeiten werden dann gruppiert und diese Gruppierungen bilden jeweils Arbeitspakete.
Aufgabe des Projektleiters ist es, diese Arbeitspakete in einem Projektstrukturplan zu visualisieren.[62]

In dem folgenden Abschnitt liegt der Fokus auf der **Qualität von Anforderungen**. Denn selbst bei der Anforderungsformulierung gibt es Fallen, die dazu führen können, dass sich

[58] Niebisch T. (2013), S. 14
[59] Vgl. Niebisch T. (2013), S. 9 ff
[60] Vgl. Booch et al. (2006), S.121
[61] Vgl. Rupp, C./ Queins, S. (2012), S. 241
[62] Vgl. Niebisch (2013), S. 11 ff

Fehler auf den gesamten Projektverlauf auswirken. Für die Formulierung von Software-Anforderungen gibt es einen vorgeschlagenen Standard: IEEE 830-1998.
Die beiden bedeutendsten Qualitätsmerkmale sind Prüfbarkeit und Eindeutigkeit.
Um zu prüfen, ob die Anforderungen erfüllt sind, ist es wichtig, auf definierte Parameter zurückzugreifen und so die Leistung subjektiv abnehmen zu können.
Das Merkmal der Eindeutigkeit greift den bereits erwähnten Punkt auf, dass keine Interpretationsmöglichkeiten für die Formulierung der Anforderungen möglich sein dürfen. Es empfiehlt sich daher, die Anforderungen in aktiver Form zu verfassen und nicht in passiver. Denn ein im Aktiv formulierter Satz beinhaltet einen Akteur. Dies ist wichtig in Bezug auf Berechtigungen und Rollen. Bei einem Satz, der im Passiv formuliert ist, bleibt der Akteur unbekannt. Bei der Formulierung muss sich ins Bewusstsein gerufen werden, dass alle Stakeholder diese Formulierung verstehen müssen.
Ein weiteres Merkmal für die Qualität einer Anforderung ist die Vollständigkeit. Bei der Recherche von Anforderungen werden dem Anforderungsmanager von Fachbereichen und Stakeholdern Anforderungen genannt. Neben diesen Anforderungen gibt es jedoch auch noch unausgesprochene Anforderungen, von denen der Mitarbeiter des Fachbereichs oder der Stakeholder ausgeht, dass diese selbstverständlich sind und es daher keiner Erwähnung bedarf. Wird beispielsweise ein bestehender Prozess geändert und nur teilweise angepasst, so werden häufig nur die vorgesehen Änderungen genannt und der unveränderte Teil des Prozesses wird außer Acht gelassen und nicht beschrieben. Eine vollständige Dokumentation wird vor allem dann wichtig, wenn nicht sicher ist, dass die nicht erwähnten Informationen jedem Projektbeteiligten bekannt sind. Zur Vollständigkeit gehört auch ein Gesamtbild aller Anforderungen, in dem die angeforderten Funktionalitäten ausführlich beschrieben werden, ohne Lücken für Interpretationen.
Priorität ist ein weiteres Merkmal und ist vor allem wichtig, weil die verschiedenen Anforderungen nicht alle sofort realisierbar sind. Daher ist es wichtig, Prioritäten zu bestimmen und die Anforderungen zu sortieren. Hier kann auf Formulierungen des deutschen Rechts zurückgegriffen werden. Zum Beispiel: Anforderung eins **muss** umgesetzt werden und ist daher eine zwingend erforderliche Anforderung. Anforderung zwei **sollte** umgesetzt werden und Anforderung drei **wird** umgesetzt, das heißt Anforderung drei ist eine Option für die Zukunft. Weitere Prioritäten wären beispielsweise Vorgaben durch den Gesetzgeber oder ein vorgegebener finanzieller Rahmen. [63]

Neben der Priorität als Qualitätsmerkmal ist Konsistenz ein wichtiges Stichwort. Die Anforderungen sollen in sich schlüssig sein und nicht mit anderen Anforderungen in einem Wider-

[63] Vgl. Niebisch (2013), S. 29 ff

spruch stehen. Eine qualitativ hochwertige Anforderung sollte außerdem nicht weiter unterteilbar sein innerhalb der Detailebene der Anforderungsformulierung.
Das Verwenden von Satzschablonen ist ein weiterer Faktor, der die Qualität der Anforderung beeinflusst. Durch den Einsatz von Satzschablonen kann Prosatext vermieden werden, der im Bereich des Anforderungsmanagements oft hinderlich ist. Zwar wird der Text dadurch sehr eintönig, hilft jedoch, fehlende Informationen und Details in den Anforderungen früh aufzudecken und zu umgehen.[64]

Je nach Situation kann unterschieden werden zwischen grafischen Darstellungen und natürlich sprachlicher Formulierung. Für komplexere Darstellungen ist es oft sinnvoll mit grafischen Modellen zu arbeiten, um das Gesamtbild nicht aus den Augen zu verlieren. Hier kann wieder auf die Unified Modelling Language zurückgegriffen werden.[65]

Ein Qualitätsfaktor, der nicht unterschätzt werden sollte, ist das Vermeiden von unterschiedlichen Begriffen für denselben Sachverhalt. Unterschiedliche Begriffe können schnell zu Missverständnissen mit fatalen Folgen führen. Bei der Kommunikation von Anforderungen werden häufig Informationen geliefert, deren Formulierung Absolutismen enthalten wie zum Beispiel „immer“ oder „nie“. Hier empfiehlt es sich laut Niebisch, die Aussagen zu hinterfragen und zu untersuchen, ob es nicht doch Ausnahmen gibt, die für die Lösung wichtig sind. Je mehr Qualitätsfaktoren in Betracht gezogen werden, desto besser lässt sich die Qualität der Anforderungen steuern.[66]

Je größer ein Projekt ist und je mehr Stakeholder und Projektbeteiligt involviert sind, desto komplexer werden die Anforderungen und es steigt die Anzahl an Anforderungen. Um die Arbeit und den Umgang mit den Anforderungen zu vereinfachen, wird eine Struktur, beziehungsweise eine Systematik notwendig. Deshalb legt der Anforderungsmanager eine Struktur fest, mit der sich die Anforderungen nach Art unterteilen lassen. Hier kann beispielsweise so vorgegangen werden, dass zwischen funktionalen und nicht-funktionalen Anforderungen unterschieden wird. Je nach Kategorisierung kann dann eine Notation oder Visualisierung nach bestimmten Standards und Modellen umgesetzt werden. [67]

Betrachtet man nach der Kategorisierung die Anforderungen genauer, ist festzustellen, dass sich diese mittels Attributen spezifizieren lassen. Ein Identifikator wäre beispielsweise ein mögliches Attribut; auch Herkunft in Priorität sind mögliche Attribute. Die Verwaltung dieser

[64] Vgl. Niebisch (2013), S. 36 ff
[65] Vgl. Niebisch (2013), S. 38
[66] Vgl. Niebisch (2013), S. 34 ff
[67] Vgl. Niebisch (2013), S. 42

Attribute ist einerseits mit Aufwand verbunden, erhöht jedoch die Einsicht in die Anforderungen.[68]

Ein weiterer wichtiger Erfolgsfaktor im Bereich des Anforderungsmanagements ist das Anlegen und Verwalten eines Glossars. Oft wird hierfür keine Notwendigkeit erkannt, da beispielsweise alle Projektbeteiligten dem gleichen Unternehmen angehören. Jedoch muss auch hier sichergestellt werden, dass alle Beteiligten die gleiche „Sprache“ sprechen und alle Begriffe eindeutig sind, um Missverständnisse nicht zu erlauben.[69]

Es sollte an alle an dem Projekt beteiligten Personen kommuniziert werden, wo und wie sie auf das Glossar zugreifen können. Die Verantwortung sowie die Erhaltung des Glossars liegen in der Hand einer einzelnen Person. Bei der Verwendung eines Glossars sollte sich bewusst gemacht werden, dass es Einträge darin gibt, die sich im Laufe der Zeit ändern werden oder ergänzt werden müssen. Wie diese Änderungen und Ergänzungen durchgeführt werden, muss zu Beginn festgelegt werden. Ebenso sollte zu Beginn definiert sein, wie Änderungen am Glossar an die Projektbeteiligten kommuniziert werden.
Dieser erste große Abschnitt wird mit der Ausarbeitung eines Anforderungsmanagement-Plans abgeschlossen. Ziel ist es, mit allen Stakeholdern, die am Anforderungsmanagement beteiligt sind, transparente Abmachungen zu treffen und Vorgaben zu dokumentieren. Dieser ausgearbeitete Plan dient dann zur Orientierung während der Anforderungsmanagement-Phase. Für einen solchen Plan gibt es keine Standards oder vorgegebenen Inhalte.[70]

Nach Abschluss dieses Punktes ist die nächste Herausforderung innerhalb des Anforderungsmanagements die Erarbeitung des Ausgangspunktes. Damit sollen Inhalte bestimmt werden und Grenzen definiert werden. Thomas Niebisch erstellt hierfür einen Fragenkatalog (siehe Tabelle 2), dessen Beantwortung dabei unterstützt, den Ausganspunkt zu definieren.[71]

[68] Vgl. Niebisch (2013), S. 42
[69] Vgl. Niebisch (2013), S. 43
[70] Vgl. Niebisch (2013), S. 45 ff
[71] Vgl. Niebisch (2013), S. 53

#	Frage
1	Welche Gründe sprechen für das Projekt?
2	Was sind die eigentlichen Ziele, die der Auftraggeber verfolgt?
3	Was ist Bestandteil des Projektes und was nicht bzw. wo verlaufen die Projekt-grenzen und wo liegen sie derzeit im Nebel?
4	Unterwelchen Annahmen, Einschränkungen und Rahmenbedingungen findet das Projekt statt?
5	Welche Personen oder Gruppen sollen oder müssen im Projekt berücksichtigt werden?
6	Wer kann wichtige Informationen oder Anforderungen nennen?

Tabelle 2: Fragenkatalog zur Erarbeitung des Ausgangspunktes[72]

In den Fragen werden unterschiedliche Perspektiven aufgegriffen und sie untersuchen jedes neue Projekt auf deren Einmaligkeit.

Auf dem Weg zur Findung des Ausgangspunktes stellt sich eine weitere Frage: Wer will was erreichen? Es ist für das Anforderungsmanagement notwendig herauszuarbeiten, welche Beweggründe die jeweiligen Projektbeteiligten haben. Um nun als Anforderungsmanager den richtigen Weg einzuschlagen, ist es wichtig, ein klares Ziel vor Augen zu haben. Der Zielfindungsprozess ist üblicherweise Teil des Projektmanagements. Diese Ergebnisse werden dann innerhalb des Anforderungsmanagements weiter verarbeitet und es findet eine genaue Abgrenzung statt, die festlegt, was Teil des Projektes ist und was nicht. Das heißt, es wird festgelegt, was in Hinblick auf die Ziele mit in das Konzept aufgenommen wird.[73]

Teil der Ausgangspunktfindung ist die Stakeholderanalyse. In diesem Schritt werden für die verschiedenen Anforderungen Inhaber definiert. Stakeholder werden in dem Gabler Wirtschaftslexikon folgendermaßen definiert: „Anspruchsgruppen sind alle internen und externen Personengruppen, die von den unternehmerischen Tätigkeiten gegenwärtig oder in Zukunft direkt oder indirekt betroffen sind.“[74]

Für die Stakeholderanalyse empfiehlt es sich ebenso ein Diagramm zur Veranschaulichung zu erstellen. So können Lücken leichter aufgedeckt werden und schneller behoben werden.

[72] Vgl. Niebisch (2013), S. 53 ff
[73] Vgl. Niebisch (2013), S. 59 ff
[74] Vgl. Thommen, J.-P. (2015), S. 1

Abschließend zur Findung des Ausgangspunktes werden Annahmen und Rahmenbedingungen gesammelt und dokumentiert. So wird mehr Klarheit darüber gewonnen welche Faktoren zur berücksichtigen sind.[75]

Die nächste Phase ist nun die **Erhebung der Anforderungen**. Hierzu gibt es mehrere mögliche Methoden, wie zum Beispiel Brainstorming, Interview, Fragebogen, Dokumentenanalyse oder Workshops. Je nach Anwendungsfall wird hier die passende Methode gewählt.[76]

Für den Anwendungsfall, der in dieser Arbeit aufgegriffen wird, wird die Methode Interview ausgewählt.

Zur Vorbereitung der Erhebung müssen zunächst die bereits erwähnten Anforderungsarten definiert werden, um den Erhebungen eine transparente Struktur zu verleihen. Ist dies geschehen, wird die Erhebungsmethode vorbereitet. In Bezug auf das Interview wurden Fragen gesammelt, die dem Team helfen ein besseres Verständnis zu erhalten. Wichtiger Leitsatz für diese Methode ist „Alles hinterfragen!“. So können Missverständnisse von Anfang an umgangen werden.[77]

Abschließend zum Thema Anforderungsmanagement wendet sich die Arbeit der Dokumentation von Anforderungen zu.

Für die Darstellung gibt es viele Wege, sowohl bildhaft, als auch natürlichsprachlich, als auch genormt beziehungsweise nicht genormt. Sinnvoll ist es, einen Mix bestehend aus diversen Notationen zu verwenden. Die Notation muss den Anforderungen angemessen sein und für die betroffenen Stakeholder bekannt und verständlich sein.[78]

Nach der Erhebung der Anforderungen ist der nächste Schritt innerhalb eine Kontextanalyse zu reflektieren, was die Erwartungshaltung der Beteiligten gegenüber dem einzuführenden Tool ist. Mit den Ergebnissen aus der Erhebung werden dann die Prozesse erneut beschrieben und dargestellt. Hierfür gilt folgender Grundsatz: „Modelle sind per Definition gegenüber der Realität unvollständig – so auch Prozessmodelle. Abweichungen vom Standard oder Verzicht auf Informationen können vorteilhaft sein – das ist im Einzelfall zu prüfen.“[79]
Es ist sinnvoll, zwischen manuellen Prozessschritten, IT-gestützten Prozessschritten und automatisierten Prozessschritten zu unterscheiden.

[75] Vgl. Niebisch (2013), S. 53 ff
[76] Vgl. Niebisch (2013), S. 73
[77] Vgl. Niebisch (2013), S. 84 ff
[78] Vgl. Niebisch (2013), S. 92 ff
[79] Niebisch (2013), S.100

Vorgegebene Regeln und Anordnungen werden bestenfalls separat von den Prozessen abgebildet, um die Lesbarkeit einfach zu halten. Handelt es sich im Anwendungsfall um sehr komplexe Regelwerke, kann eine Entscheidungstabelle als Hilfsmittel herangezogen werden. Werden Anforderungen in der natürlichen Sprache formuliert, empfiehlt sich der Einsatz von Satzschablonen für saubere Formulierungen.[80]

Abbildung 5 zeigt wie eine solche Satzschablone aussehen kann. Hervorzuheben ist, dass es in dieser Darstellungsart einen Akteur gibt, was zur Folge hat, dass die Anforderungen aktiv und nicht passiv formuliert werden.

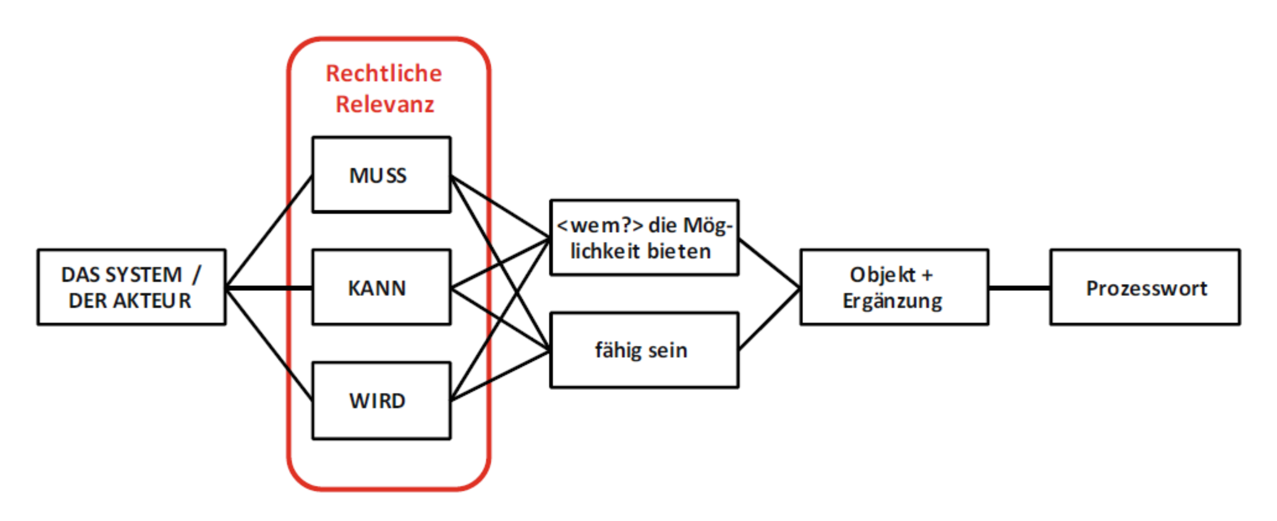


Abb. 5: Satzschablone ohne Bedingungen[81]

Durch das Arbeiten mit Satzschablonen können Qualitätsanforderungen leichter sichergestellt werden und sprachlich bedingte Missverständnisse können vermieden werden.

Für den Umgang mit Daten und deren Strukturierung sind Klassendiagramme ein hilfreiches Vorgehen. Sie unterstützen den Anforderungsmanager beim „[…]Sammeln von Dingen (Klassen), deren Merkmalen (Attributen, Eigenschaften) und Operationen sowie den Beziehungen zwischen Dingen (Klassen).“[82]

Darauf aufbauend werden Zustandsdiagramme innerhalb des Anforderungsmanagements erstellt. In einem Zustandsdiagramm werden die verschiedenen Zustände verschiedener Klassen in Bezug auf eines ihrer Attribute dargestellt.

Bereits im Kontextdiagramm werden Systemgrenzen definiert und nun wieder aufgegriffen, um sogenannte Umsysteme zu integrieren und Schnittstellen zu definieren. Eine Darstellung kann sowohl über Domänen-Modelle als auch über Choreographie-Diagramme erfolgen.

[80] Vgl. Niebisch (2013), S. 105 ff

[81] Enthalten in: Rupp, C. et al (2009), S. 162

[82] Niebisch (2013), S. 111

Beide Modelle sind nicht in der Lage, alle Anforderungen an die Schnittstellen komplett abzubilden.[83]

Für den Umgang mit Berechtigungen kann beispielsweise der CRUD-Ansatz verwendet werden. CRUD steht für:

- „Create: Datensatz anlegen
- Read: Datensatz abfragen
- Update: Datensatz ändern
- Delete: Datensatz löschen“[84]

Dahinter verbirgt sich, dass ein entwicklungsnahes Modell, in dem die Rechte, die bestimmte Nutzergruppen an bestimmten Daten haben, aufgelistet werden.

Für die Dokumentation der Anforderungen für Nutzeroberflächen wird häufig mit Dialogflüssen gearbeitet. Sie ermöglichen es, notwendige Masken und Dialoge organisiert darzustellen und Zusammenhänge, so wie Steuerelemente abzubilden.

Die genannten Schritte helfen dem Anforderungsmanager bei einer strukturierten Dokumentation der Anforderungen und zeigen verschiedene Werkzeuge und deren Einsatzmöglichkeiten auf.[85]

3 IST-SOLL-Analyse

In der IST-SOLL-Analyse wird zunächst das aktuell verwendete Dokumentationssystem der Anwendungsentwicklung der .Versicherung untersucht, um auf Basis der Kritik, Anforderungen an das neue Tool bestimmen zu können, die in Form eines Kriterienkatalogs dargestellt werden.

3.1 IST-Analyse des aktuell verwendeten Tools in der .Versicherung

Die Mitarbeiter der Anwendungsentwicklung der .Versicherung benutzen zurzeit eine abteilungsinterne Notesdatenbank. Aus dem Ausschreiben der .Versicherung , das die Anfrage nach einem neuen Dokumentationssystem für die Mitarbeiter der Anwendungsentwicklung enthält, konnten keine Eigenschaften des aktuell genutzten Dokumentationssystems abgelei-

[83] Vgl. Niebisch (2013) S. 71 ff
[84] Niebisch (2013), S. 122
[85] Vgl. Niebisch 2013), S. 121 ff

tet werden. Um eine detaillierte Beschreibung der Charakteristika des aktuellen Tools auf Basis einer IST-Analyse durchführen zu können, wurde das Gespräch mit einem Repräsentanten aus der Anwendungsentwicklung der .Versicherung gesucht. Mit Hilfe schriftlicher sowie verbaler Interviewverfahren konnten einige Merkmale des aktuell genutzten Tools identifiziert werden. Diese sind in den folgenden Paragraphen aufgeführt.

Die Notesdatenbank ist ein Ablagesystem, um Dateien ablegen, ändern und wieder auf sie zugreifen zu können. Momentan arbeiten ungefähr 75 Mitarbeiter aus der Anwendungsentwicklung mit der Notesdatenbank. Im Systemspeicher befinden sich derzeit etwa 900 Dateien, diese sind hauptsächlich Bilddateien und Textdokumente, die soziale Informationen und Best Practices enthalten. Das aktuelle System wurde vor fünf Jahren eingeführt und löste eine sehr unstrukturierte Diskussions-Notesdatenbank ab.
Das Tool bietet die Möglichkeit, diverse Kategorien flexibel anzulegen, um Dateien inhaltlich nach den angelegten Kategorien ordnen zu können. Die Flexibilität dieser Option bezieht sich auf das Bearbeiten der Kategorien, denn diese können gelöscht, hinzugefügt oder geändert werden.
Des Weiteren verfügt das Ablagesystem über einen Redaktionsprozess. Die Einbindung des Redaktionsprozesses verfolgt das Ziel, dass ältere Dokumente, die nicht mehr genutzt werden, nicht den Systemspeicher unnötig belegen. Die Prozedur des Redaktionsprozesses sieht vor, dass jedem hochgeladenem Dokument eine befristete Existenz von zwei Jahren zugewiesen wird. Bei Erreichen dieser Frist, erhält der verantwortliche User, der das Dokument hochgeladen hat, der sogenannte Autor, eine Benachrichtigung und kann dann dementsprechend reagieren: das betroffene Dokument kann, wie durch die Einbindung des Redaktionsprozesses vorgesehen, gelöscht werden oder, im Falle von Unverzichtbarkeit auf das Dokument, archiviert werden, der Lebenszyklus wird also verlängert.
Zusätzlich bietet die Notesdatenbank vorkonfigurierte Methoden zur Fehlerbehandlung. In Fällen von Systemstörungen kann der Betrieb, dieser besteht aus Rechenzentrum und Netzwerken, Benachrichtigungen per E-Mail versenden, die in der Datenbank in einer speziellen Darstellung angezeigt werden.[86]

3.2 Kritik an aktuell verwendetem Ablagesystem

Die Notesdatenbank, die die Mitarbeiter aus der Anwendungsentwicklung der .Versicherung aktuell als Dokumentationsplattform nutzt, trifft nicht die Vorstellungen des Großteils der User. Im Zuge von mehreren Interviewverfahren mit einem Repräsentanten aus der Anwen-

[86] Vgl. Ade, U. (2015)

dungsentwicklung konnten die größten Kritikpunkte an das momentan verwendete Dokusystem identifiziert werden.

Die betroffenen Mitarbeiter nennen diverse Probleme und Mängel, die aus dem Umgang mit dem Tool resultieren.
Das aktuell verwendete Tool entspricht nicht den Benutzeranforderungen der Mitarbeiter aus der Anwendungsentwicklung der .Versicherung . Vielmehr beeinträchtigt das aktuelle Ablagesystem besagte Mitarbeiter bei der Bewältigung täglich anfallender Arbeit. Das aus der Nutzung mit dem Ablagesystem resultierende Hauptproblem ist die antiquierte Benutzeroberfläche. Aus diesem Grund stuft der Großteil der Mitarbeiter aus der Anwendungsentwicklung der .Versicherung das Handling des Systems als nicht benutzerfreundlich ein und berichtet, dass die Benutzung des aktuellen Tools die Bewältigung der täglichen Arbeit erschwert.
Neben den Problemen, die ihren Ursprung in der nicht benutzerfreundlichen graphischen Oberfläche haben, konnten weitere Kritikpunkte erfasst werden.
Zum einen bietet das System kaum bis keine Möglichkeiten für Layoutanpassungen und Formatierungen. Die Bemängelung fehlender Layoutanpassungen bezieht sich vor allem auf Designänderungen. Viele User wünschen sich ein ansprechenderes Design, das auch bei Bedarf wieder geändert werden kann.
Des Weiteren entspricht die Suchfunktion nicht der Benutzererwartung. Die Such-Logik der Notesdatenbank ist nach Angaben des Repräsentanten der Anwendungsentwicklung sehr unterschiedlich im Vergleich zu der bekannten Such-Logik aus der Google-Suchmaschine. Die nicht nachvollziehbare Struktur der Such-Logik der Notesdatenbank schlägt sich vor allem im Zeitaufwand negativ nieder, denn das Suchen einzelner Dokumente dauert. Es zeigt sich, dass der erhöhte Zeitaufwand zu Unmut bei den Benutzern führt und sie bei dem Erledigen ihrer Aufgaben deutlich behindert und einschränkt.
Auch die Erfassung der Dokumente, das Hochladen der Dokumente stellt für viele Benutzer eine Hürde im Umgang mit dem Tool dar. Die Erfassung kann einerseits über das Einbinden von Anhängen und andererseits über die Kurzbefehle „copy“ und „paste“ in das eigentliche System stattfinden. Beide Vorgehensweisen sind laut Aussage des Interviewpartners kompliziert. Im Falle der Einbindung von Daten über Anhänge ist die Öffnung der Dokumente sehr kompliziert. Die Dateien können nicht via Doppelmausklick geöffnet werden, da für die Dateiöffnung kein Default-Programm festgelegt werden kann. Die Dokumente können lediglich über einen Rechtsklick mit der Maus über die Auswahl „Mit Word öffnen“ geöffnet werden. Werden die Dateien im System selbst gespeichert, über die Kurzbefehle „copy“ und „paste“, gehen bis zu 80% der Wordformatierung verloren. Auch die Verlinkung von Dokumenten ist zwar möglich, jedoch ebenfalls sehr kompliziert.[87] [88]

[87] Vgl. Ade, U. (2014)

Es lässt sich zusammenfassen, dass das aktuell genutzte System zur Dokumentenablage Inakzeptanz seitens der User hervorruft. In Rückblick auf Kapitel 2.1.2 (Wissensmanagement) kann festegehalten werden, dass das aktuell betriebene Wissensmanagement in der Anwendungsentwicklung der .Versicherung nicht den Elementen des in Kapitel 2.1.2 erläutertern TOM-Modells entspricht.

3.3 Anforderungen an das neue Tool

Die Ergebnisse der durchgeführten IST-Analyse und die identifizierten Kritikpunkte an dem aktuellen Dokumentationssystem der Anwendungsentwicklung der .Versicherung zeigen in Bezug auf Wissensbewahrung drei gravierende Probleme auf. Es kann festgehalten werden, dass zu viel Wissen verloren geht und dass Wissen nicht aktualisiert wird. Aus diesen beiden Erkenntnissen wird gefolgert, dass Wissen nicht effektiv genutzt wird.

Für die Arbeit der Mitarbeiter aus der Anwendungsentwicklung der .Versicherung ist es aus diesem Grund essentiell, ein neues Dokumentensystem zu implementieren, welches den Bedarfsanforderungen der User entspricht.

Auf Basis bereits bestehender Informationen und auf Basis von Interviews kann ein Kriterienkatalog, welcher das Entscheidungsinstrument für das neue Dokumentationssystem ist, definiert werden. Mit Ausschreibung der Anfrage nach einem neuen Doku-Ablagesystem gab die .Versicherung einige Grundanforderungen an das neue Tool bekannt.

Die Hauptgrundanforderung ist in erster Linie, dass das alternative Dokumentationssystem ein Open Source Produkt und eine Standardtechnologie sein soll. Open Source Software muss unter einer Lizenz stehen, die offiziell von der Open Source Initiative (OSI) anerkannt wird. Demnach muss Open Source Software laut der offiziellen Website der Open Source Initiative „opensource.org“ zehn Kriterien erfüllen. Die zehn Kriterien lauten:

1. Freie Umverteilung der Lizenzen („Free Redistribution“)
2. Bereitstellung und Erlaubnis zu Umverteilung des Source Codes („Source Code“)
3. Erlaubnis zu Sourcecodeänderungen („Derived Works“)
4. Abgeleitete Software bzw. Source Code Änderungen dürfen nicht verteilt werden und unter Umständen umbenannt werden („Integrity of The Author’s Source Code“)
5. Die Lizenz darf sich nicht gegen einzelne Personen oder Personengruppen wenden („No Discrimination Against Persons or Groups“)
6. Keine Einschränkungen im Anwendungsbereich („No Discrimination Against Fields of Endeavor“)
7. Mit Erhalt der Software muss die Lizenz mitgeliefert werden („Distribution of License“)

[88] Vgl. Ade, U. (2015)

8. Die Lizenz der Software darf sich nicht auf eine spezielles Produkt beziehen („License Must Not Be Specified to a Product")
9. Einzelne Lizenzen dürfen keine andere Software des Softwarebündels einschränken („License Must Not Restrict Other Software")
10. Individuelle Technologien dürfen nicht die Grundlage für Lizenzen sein („License Must Be Technology-Neutral")[89]

Nach Rücksprache mit dem Interviewpartner aus der Anwendungsentwicklung wurden, bezogen auf das Kriterium „Open Source", Informationen hinzugefügt. Es wurde erklärt, dass, sofern das empfohlene neue Dokusystem nicht kostenlos sein sollte, jedoch alle anderen Anforderungskriterien trifft, ein Budget durchaus angefragt werden kann.[90]
Andere Kriterien, die sich aus dem Anforderungskatalog ableiten lassen, sind zum einen der Funktionsumfang der Software, was ebenfalls die Grundlage für weitere Gestaltungsmöglichkeiten legt, sowie eine Suchfunktion, die die Logikerwartungen der Benutzer trifft, und der Umgang mit Anhängen, was in diesem Fall Bilder und Officedokumente, größtenteils Textdokumente, sind.
Zusätzliche Kriterien, die aus der Ausschreibung zu entnehmen sind, sind die Möglichkeit der Einrichtung eines Autorensystems und eine Rechte- und Zugriffsverwaltung sowie die technische Datenverwaltung und Größenbeschränkungen.

Diese oben beschriebenen Anforderungen sind im Hinblick auf Marktuntersuchungen nach einem geeigneten Produkt noch nicht spezifisch genug. Aus diesem Grund wurden Interviews mit einem stellvertretenden Mitarbeiter aus der Anwendungsentwicklung der .Versicherung geführt. Im Nachgang dieser Interviews konnten bereits vorhandene Kriterien näher spezifiziert sowie weitere Kriterien identifiziert werden.
Zunächst einmal wurde geklärt, was mit Standardtechnologien im Detail gemeint ist. Eine Standardtechnologie, so wie sie sich die Mitarbeiter aus der Anwendungsentwicklung der .Versicherung wünschen, sollte offene Schnittstellen bereithalten, was vor allem System-Komptabilität, wie beispielsweise die Komptabilität mit Microsoft Office Dokumenten, bedeutet. In Bezug auf bereits identifizierte Anforderungspunkte wurde hinzugefügt, dass die Bildung von Benutzergruppen, die unterschiedliche Zugriffsrechte, wie ‚Lesen', ‚Bearbeiten' und weitere, nicht relevant für die Arbeit der Mitarbeiter aus der Anwendungsentwicklung der .Versicherung sei.[91]
Ein Autorensystem, also eine Verwaltung, die jedem Benutzer anzeigt, welche Dokumente von welchem Benutzer (oder Autor) hochgeladen worden sind und wie viele und welche Do-

[89] Vgl. Hugos, M./Hulitzky D. (2011), S. 55ff.
[90] Vgl. Ade, U. (2014)
[91] Vgl. Ade, U. (2014)

kumente ein spezieller Benutzer insgesamt hochgeladen hat, ist dagegen jedoch wichtig. Des Weiteren wurde das Thema ‚Größenbeschränkungen' im Hinblick auf Useranzahl und im Hinblick auf abzulegende Datenmengen, näher analysiert. Aktuell sind lediglich ungefähr 70 Mitarbeiter in der Anwendungsentwicklung der .Versicherung tätig, jedoch ist eine zukünftige Anzahl zwischen 100 und 150 Usern realistisch. Die aktuelle Anzahl an abgelegten Dokumenten, diese sind hauptsächlich Bilder, Textdokumente und PDF-Dateien, beträgt ungefähr 900, realistisch ist jedoch eine zukünftige Anzahl von nicht mehr als 20.000 Dokumenten.[92]

Im Hinblick auf die zu erwartende größer werdende Dokumentenanzahl ist es den Mitarbeitern aus der Anwendungsentwicklung der .Versicherung sehr wichtig, dass die neue Dokumentationsablage ein Versionierungssystem bereithält.

Ein weiteres äußerst wichtiges Anliegen der Benutzer ist, dass die Navigation sehr einfach ist und dass das neue System keine großen Hierarchiestrukturen bereithält, also dass der Zugriff auf Dokumente durch einfache Klicks stattfindet und dass das Klicken durch mehrere Ordner, um auf das gewünschte Dokument zugreifen zu können, vermieden wird.

Ebenfalls ein Kriterium, das die Benutzer als notwendig bewerten, ist die technische Umsetzung der Dateiablage. Sämtliche Dokumente sollen durchsuchbar sein. Entweder können die Dateien im System selbst abgelegt werden und durch Icon-Verlinkungen kann auf das Dokument extra zugegriffen werden. Eine andere Lösung ist das Hochladen von Anhängen, um hier gewährleisten zu können, dass diese durchsuchbar sind, sollten die Dokumente durch spezifische Schlagwörter, sogenannte Tags, charakterisiert werden.

Des Weiteren wurden noch drei andere Kriterien genannt: eine einfache Implementierung, die Einsetzbarkeit und die Verbreitung bzw. der Reifegrad.

Für die Anwendungsentwickler ist es wichtig, dass die Implementierung des Systems simpel gehalten ist und nicht mit einem großen Mehraufwand in Form von langen Systemtestingphasen oder dem Erwerb von mehr Servern verbunden ist.

Das Kriterium Einsetzbarkeit bedeutet, ob das Dokusystem auch auf anderen Endgeräten, wie beispielsweise Smartphones oder Tablets, verfügbar ist.[93]

Auch Informationen über den Reifegrad bzw. über die Verbreitung des Systems ist wichtig für die Entscheidungsfindung: für Systeme, die noch nicht weit ausgereift sind oder noch keine Anwendung finden, gibt es häufig kaum bis keinen Support.

[92] Vgl. Ade, U. (2014)
[93] Vgl. Ade, U. (2014)

3.4 Erstellung eines Kriterienkatalogs

Nach Betrachtung aller Anforderungen an das neue Dokusystem und in Rücksprache mit dem Interviewpartner aus der Anwendungsentwicklung der .Versicherung , wurde ein Kriterienkatalog entwickelt. Bei der Erstellung des Kriterienkatalogs wurden die Theorien des Anforderungsmanagements aufgegriffen und es wurde mit Satzschablonen bei der Anforderungsformulierung gearbeitet. Der Kriterienkatalog, siehe Figur 1, beinhaltet sämtliche Anforderungen, die das Ablösungstool für die Dokumentation erfüllen sollte, nach Wichtigkeit und Dringlichkeit anhand der Prioritätsstufen „1“ (unverzichtbar) bis „5“ (am wenigsten wichtig) bewertet.

Priorität	Kriterium	Anforderung	Tool 1	Tool 2	Tool 3	Tool 4	Tool 5
1	Open Source	Das System muss eine Open Source Software sein.					
1	Suchfunktion	Der Nutzer muss die Möglichkeit haben das gesamte System nach Schlagworten durchsuchen zu können.					
1	Einfache Navigation	Das System muss dem Nutzer eine einfache Systemnavigation ohne Hierarchiestrukturen bieten.					
1	Ablage von Dateien	Das System muss dem Nutzer die Möglichkeit bieten Dateien hochzuladen.					
2	70-150 User	Das System sollte für mindestens 150 Benutzer ausgelegt sein.					
2	Versionierung	Das System sollte dem Nutzer die Möglichkeit bieten bei der Verwaltung von Dateien eine Versionierung einzusehen.					
2	Speicher Datenmengen <20.000	Das System sollte die Möglichkeit bieten bis zu 20.000 Dokumente hochzuladen.					
3	Verbreitung/Reife	Das System sollte fähig sein Referenzen aufzuweisen, die die Verbreitung und Reife des Systems auszeichnen.					
3	einfache Implementierung	Das System sollte es ermöglichen eine Implementierung nach Standards durchzuführen.					
3	Datengröße	Das System sollte dem Nutzer die Möglichkeit bieten Dateien der Größe von bis zu 5 MB hochzuladen.					
4	Handling von Anhängen	Die Software sollte den Nutzern die Möglichkeit bieten Dateien direkt abzulegen oder hinter Icon-Verlinkungen zu hinterlegen.					
4	Lizenzmodell	Das System sollte einem OpenSource Lizenzmodell unterliegen.					
4	Berechtigungssystem	Das System soll es ermöglichen Benutzergruppen zu bilden und Benutzern unterschiedliche Zugriffsrechte einzurichten.					
5	Gestaltungsmöglichkeiten	Das System wird der Halleschen KVaG die Möglichkeit bieten über zusätzliche Funktionen die Gestaltungsmöglichkeiten zu erweitern.					
5	Einsetzbarkeit	Das System wird für den Benutzer auch auf anderen Endgeräten, wie z.B. Smartphones oder Tablets verfügbar sein.					
5	Autorensystem	Das System wird dem Benutzer die Möglichkeit bieten auf ein Autorensystem zugreifen zu können und dieses zu verwalten.					

Tabelle 3: Kriterienkatalog

Die Top eins Prioritäten für die Auswahl des neuen Dokusystems sind die Merkmale Open Source und eine einfache Suchfunktion, die sämtliche Dokumente nach Schlagwörtern durcharbeitet. Bedingt durch ein ansprechendes Layout sind eine simple Navigation und die Möglichkeit der Dateiablage „unverzichtbar“.

Als ebenfalls „sehr wichtig“, jedoch nachrangig gegenüber den oben genannten Top eins Prioritäten, sind die zulässige Anzahl an Usern und an Datenmengen sowie die Versionierung.

Die drei Kriterien Reifegrad des Systems, eine einfache Implementierung sowie der verfügbare Speicherplatz haben die Priorität drei „wichtig“.
Die vierte Gruppe bilden die „weniger wichtigen“(Priorität vier) Kriterien: Das Handling von Anhängen, das Lizenzmodell sowie die Implementierung eines Berechtigungssystems.
Gestaltungsmöglichkeiten des Layouts sowie die Verfügbarkeit des Systems auf mobilen Endgeräten und die Erfassung eines Autorensystems befindet die Kontaktperson aus der Anwendungsentwicklung der .Versicherung als „am wenigsten wichtig“ (Priorität fünf).[94]

4 Marktanalyse

In dem folgenden Kapitel wird das Vorgehen der Marktanalyse erläutert und erste Marktuntersuchungen werden dokumentiert.

4.1 Methodisches Vorgehen

Im Zuge der IST-SOLL-Analyse aus dem vorangegangenen Kapitel ist ein Kriterienkatalog erstellt worden, mit dessen Hilfe der Markt nach Produkten gescannt werden soll, die die definierten Kriterien erfüllen und als Dokumentenablagesystem für die Mitarbeiter der Anwendungsentwicklung der .Versicherung geeignet ist.

Nach ersten Recherchen hat sich herausgestellt, dass der Markt an Open Source Produkten, die als Dokumentenverwaltungssystem fungieren, sehr groß ist. Um aus der Menge des Angebots eine realistische Anzahl an zu analysierende Systeme identifizieren zu können, ist der Markt in mehreren Durchgängen gefiltert worden. Aus diesem Grund haben alle Autoren dieser Arbeit die Funktion der Marktforschung übernommen.

In der ersten Markt-Recherche hat Jeder der sechs Marktforscher eine eigene Marktanalyse durchgeführt. Ziel des ersten Vorgangs war es, eine Liste von mindestens fünf Produkten, die sich als Dokusystem für die Anwendungsentwicklung der .Versicherung eignen, zu erstellen. Die einzelnen Ergebnisse aller sechs Marktforscher sind in Kapitel 4.2 (Ergebnisse der Marktanalyse) aufgeführt. Insgesamt sind 22 potenzielle Systeme identifiziert worden. Die erste Marktanalyse wurde auf Basis zuvor vereinbarter K.O.-Kriterien (engl. Knockout) durchgeführt. Die zuvor abgestimmten K.O.-Kriterien sind alle Kriterien aus dem Kriterienkatalog mit der Priorität eins. In Bezug auf den Kriterienkatalog sind die K.O.-Kriterien Open Source, eine einfache Navigation und die Ablage von Dokumenten. Das vierte Kriterium der Gruppe, eine einfache Suchfunktion, wurde in der ersten Marktanalyse nicht berücksichtigt.

[94] Vgl. Ade, U. (2015)

Dieses Kriterium wurde außen vor gelassen, da es sehr schwierig ist, diese Eigenschaft anhand von Websites und Foren zu untersuchen.

In dem zweiten Durchgang der Marktanalyse wurde die gesamte Liste an potenziellen Produkten auf fünf zu analysierende Produkte reduziert. Die Reduzierung der Liste kann in Kapitel 4.4 (Die Top fünf Tools) und in Kapitel 4.5 (Nachbereitung der Top fünf Tools: Finale Liste) eingesehen werden. Die Verkürzung der Liste auf fünf Produkte ist durchgeführt worden, indem jeder Marktforscher das Favoriten-Tool aus der eigenen Recherche gewählt hat. Die Favorisierung der Systeme basiert auf verschiedene Gründen, diese sind in Kapitel 4.4 und ergänzend in Kapitel 4.5 näher erläutert.

Die fünf Dokusysteme werden auf Basis des Kriterienkatalogs aus Kapitel 4.4 (Erstellung eines Kriterienkatalogs) einzeln analysiert. Als Analyse-Referenzen dienen die Websites der jeweiligen Tools und ergänzend dazu, werden Interviews mit den Herstellern der Tools geführt.

Nachdem alle fünf Tools der finalen Liste einzeln analysiert wurden, werden die Ergebnisse in einer Übersicht zusammengetragen, siehe Kapitel 5.6. Die Übersicht dient dazu, die Auswahl näher einzugrenzen, denn Ziel ist es, der .Versicherung ein Dokumentationssystem zu empfehlen.

Wurden Favoriten anhand der Übersicht bestimmt, werden diese einer Nutzwertanalyse unterzogen. Nutzwertanalysen dienen als Instrument der Entscheidungsunterstützung und helfen dem Entscheider eine Entscheidung nach ihren oder seinen Vorstellungen zu treffen. Das bedeutet konkret, dass die einzelnen Alternativen nach den Vorstellungen des Entscheiders seinen Vorstellungen nach in Form von Nutzwerten geordnet werden.[95]
Bei der Erstellung einer Nutzwertanalyse wird folgendermaßen vorgegangen: Die einzelnen Kriterien werden zunächst nach den Vorstellungen des Entscheiders priorisiert. Hierbei ist zu beachten, dass die wichtigsten Kriterien mit hohen Zahlenwerten belegt werden, das heißt, dass die Prioritäten absteigend gewählt werden. Nach erfolgreicher Priorisierung werden die einzelnen Alternativen in Bezug zu den Kriterien, unabhängig von der zuvor gewählten Priorisierung, bewertet. Am Ende der Bewertung werden die einzelnen Werte mit den Priorisierungswerten multipliziert und diese werden dann pro Alternative aufaddiert.[96]

Die Ergebnisse der Nutzwertanalyse unterzogenen Favoriten können in Kapitel 5.7 (Finale Empfehlung mit Hilfe einer Nutzwertanalyse) eingesehen werden.

[95] Vgl. Mehlan, A. (2007), S.55f.
[96] Vgl. Weber, J. (2007), S.2f.

4.2 Ergebnisse der Marktanalyse

Nach erfolgreich durchgeführter Analyse der am Markt verfügbaren Produkte, sind die Ergebnisse aller sechs Marktforscher zusammengetragen worden.
Einige der als geeignet befundenen Produkte sind mehrfach genannt worden. Tools, die mehrfach in den einzelnen Ergebnissen aufgetaucht sind, sind vor allem Produkte, die einen höheren Bekanntheitsgrad haben. Beispiele für doppelt genannte Produkte sind „Google Drive" und „Dropbox".
Alle Ergebnisse aus der Marktuntersuchung sind im Folgenden aufgelistet:

- Agorum
- Alfresco
- DokuWiki
- Dropbox
- Drupal Wiki
- edx platform
- FlexWiki
- Foswiki
- Google Drive
- Huddle
- Instiki
- LetoDMS
- MediaWiki
- MindTouch
- Scribble Papers
- TikiWiki
- TreeSheets
- Twiki
- UDocs
- WikiPad
- Wix
- Wiki4enterprise
- WordPress

4.3 Die Top 5 Tools

Die Entscheidung über die Auswahl der am besten geeignetsten Tools als Dokusysteme für die Anwendungsentwicklung der .Versicherung ist auf folgende fünf Dokusysteme gefallen:

- Agorum
- DokuWiki
- MediaWiki
- Scribblepapers
- UDocs

Die Gründe für die Wahl dieser fünf Systeme sind in den folgenden Abschnitten erläutert.

„Agorum" hat eine sehr gute Internetpräsenz, sämtliche Eigenschaften des Tools sind auf der Website anschaulich dargestellt. Ein ebenfalls wichtiges Kriterium ist die Toolart bzw. Toolsorte: „Agorum" ist ein klassisches Dokumentenverwaltungssystem und kein Wiki.

„DokuWiki" ist ein Wiki-System. Es wurde in erster Linie deshalb ausgewählt, weil es auf vielen Ratingportalen im Internet als das beste und als das am häufigsten verwendete Wiki genannt wird. Zusätzlich enthält „DokuWiki" viele Eigenschaften, die den Anforderungen an das neue Tool entsprechen. „DokuWiki" stellt eine Suchfunktion sowie einen Index bereit und es gibt die Möglichkeit, das Layout des Systems individuell zu ändern. Ein Punkt, der das Tool „DokuWiki" ganz besonders für die Anwendungsentwicklung der .Versicherung interessant macht, ist, dass „DokuWiki" speziell auf Gruppen wie Entwicklerteams ausgelegt ist.[97] [98]

„Media Wiki" ist ebenfalls ein Wiki-System, für das sich das Marktforschungsteam entschieden hat, weil es viele der definierten Anforderungen erfüllt. Die Benutzung von „Media Wiki" ist an die Benutzung von der Online-Wissensdatenbank „Wikipedia" angelehnt. Durch die große Bekanntheit von Wikipedia hat „Media Wiki" den Vorteil, dass die Anwender mit der Benutzeroberfläche des Tools vertraut sein werden, was die Usability erleichtert.[99]

„Scribble Papers" wird als eine Art „Zettelkasten"[100] beschrieben. Es wurde aufgrund seiner Vielzahl an Vorteilen und Eigenschaften, die die Anforderungen an das neue Tool treffen, in

[97] Vgl DokuWiki (2015a)
[98] Vgl DokuWiki (2014a)
[99] Vgl. o.V. (2014c)
[100] o. V. (2014b)

die Top fünf Liste gewählt. Die Bereitstellung einer einfach gestalteten Suchfunktion sowie die benutzerfreundliche graphische Oberfläche des Tools sind die beiden Hauptargumente für die Auswahl dieses Tools. Weitere Vorteile von „Scribble Papers“ sind die Option der individuellen Layoutanpassung sowie die Unterstützung vieler verschiedener Dateiformate.[101]

Das Tool „UDocs“ hat durch drei große Vorteile überzeugt. Zum einen ist die graphische Benutzeroberfläche ansprechend gestaltet, was eine einfache Navigation zur Folge hat. Zum anderen bietet „UDocs“ eine Archivierungsfunktion. Der größte Vorteil, den das Tool für die Top fünf Liste der Tools qualifiziert hat, ist die Bereitstellung von zwei verschiedenen Suchfunktionen: mittels der einen Suchfunktion kann nach Schlagwörtern gesucht werden, die andere Suchfunktion kann für die Suche nach Dateien benutzt werden.[102]

4.4 Nachbereitung der Top 5 Tools: Finale Liste

Nachdem die einzelnen Tools aus der Top 5 Liste auf Basis des Kriterienkatalogs detaillierter untersucht wurden und nach Kontaktaufnahme zu den Herstellern der einzelnen Tools, ist das Marktforschungsteam zu dem Ergebnis gekommen, dass die finale Liste der Top fünf Tools für die Dokumentenablage überarbeitet werden muss.
Die beiden Dokusysteme „Scribble Papers“ sowie „Udocs“ müssen ersetzt werden.

Nachdem die Analyse der einzelnen Tools auf Basis des Kriterienkatalogs gestartet wurde, hat sich herausgestellt, dass „Scribble Papers“ kein Open Source Produkt ist. Es ist zwar kostenlos verfügbar, jedoch liegt der Quellcode nicht offen zur Verfügung. Des Weiteren hat sich ergeben, dass das Tool unpraktisch für die Zusammenarbeit der Mitarbeiter aus der Anwendungsentwicklung der .Versicherung ist. Denn es können zwar mehrere Benutzer auf das Tool zugreifen, jedoch hat nur ein Benutzer einen „Schreibezugriff“, dem Rest der Benutzer werden lediglich „Lesezugriffe“ zugeteilt. Das bedeutet konkret, dass nur ein Benutzer Dateien hochladen, ändern und entfernen kann, die verbleibenden Benutzer haben lediglich das Recht, die Datei zu lesen oder anzuschauen.[103]

Im Zuge der Analyse der einzelnen Tools musste auch das Dokusystem „UDocs“ aus der Liste der Top fünf Produkte entfernt werden. Es hat sich herausgestellt, dass „UDocs“ die Dokumentation einer speziellen Cloud Software des Herstellers „Unicon Software“ ist. „U-

[101] Vgl. Humpa, M. (o. J.)
[102] Vgl. o. V. (2013)
[103] Vgl. Hötger, J. (2015)

Docs“ stellt somit folglich nur eine Art Benutzerhandbuch für die selbstständige Fehlerlösung der Software von Seiten der Benutzers dar.[104]

Die finale Liste der Tools, die für die Anwendungsentwicklung der .Versicherung als Dokumentenablage in Frage kommen und auf Basis des in Kapitel 3.4 erstellten Kriterienkatalogs analysiert werden, um eine Empfehlung abgeben zu können, sind:

- Agorum
- Alfresco
- DokuWiki
- MediaWiki
- Wordpress

Die Begründungen für die Entscheidung für die Ersatztools „Alfresco“ und „Wordpress“ sind in den folgenden Abschnitten erläutert.

„Alfresco“ hat einen sehr hohen Bekanntheitsgrad, es ist durch viele internationale Kunden weltweit vertreten: über 11 Millionen Benutzer greifen auf „Alfresco“ zurück. Bedingt durch die weite Verbreitung von „Alfresco“ bietet der Hersteller einen sehr guten Support, der rund um die Uhr und sieben Tage die Woche läuft. Auch nach näherer Untersuchung entspricht „Alfresco“ allen Anforderungen. Ein Nachteil wäre, dass Alfresco kostenpflichtig ist.[105]

„WordPress“ ist ein Blog-System und hat eine sehr simple Benutzeroberfläche, dementsprechend ist die Usability sehr hoch. Durch die Aktivierung der Funktion „Multisite“ kann ein Netzwerk mit beliebig vielen WordPress-Instanzen eingerichtet werden. Des Weiteren gibt es unzählig viele Funktionserweiterungen in Form von Plugins, mit denen „Wordpress“ beliebig und individuell erweitert und gestaltet werden kann.[106]

5 Analyse der Tools anhand des Kriterienkatalogs

Im folgenden Kapitel werden die Top fünf Tools aus Punkt 4.4 einzeln anhand der Kriterien des Kriterienkataloges analysiert.

[104] Vgl. o. V. (2014d)
[105] Vgl. o. V. (2015a)
[106] Vg, o. V. (o. J.)

5.1 Agorum Core Open

Das Agorum Core ist ein Dokumentenverwaltungssystem der Agorum Software GmbH. Es gibt zwei Versionen, das Agorum Core Open und das Agorum Core Pro. In dieser Arbeit wird das Produkt Agroum Core Open betrachtet, da ausschließlich Open Source Programme für die .Versicherung in Frage kommen.[107] Im Folgenden wird nun das Produkt auf die zuvor definierten Kriterien untersucht.

5.1.1 Kriterien

Priorität	Kriterium	Agorum Core Open	Kommentar
1	Open Source	x	
1	Suchfunktion	x	keine Highlighting in Open Source Version
1	Einfache Navigation	x	
1	Ablage von Dateien	x	Ordner, Dokumente, E-Mails, Wikis, Foren, Termine, Benutzer, Gruppen, Eigene Objekttypen
2	70-150 User	x	
2	Versionierung	x	
2	Speicher Datenmengen <20.000	x	
3	Verbreitung/Reife	x	deutsch, 10000 Downloads, Produkt seit 2008, Unternehmen seit 1998, gute Referenzen
3	einfache Implementierung	x	
3	Datengröße	x	
4	Handling von Anhängen	?	Speicherort nicht exakt bekannt, Server ist anzunehmen, keine Vorschau
4	Lizenzmodell	x	GPL2

[107] Vgl. agorum® Software GmbH (o.J.a)

4	Berechtigungssystem	x	Gruppen, Benutzer, Access Control Lists
5	Gestaltungsmöglichkeiten	x	Ordner-basiert, eigene Objekttypen mit eigenen Attributen, Meta-Daten, Integration eigener Masken
5	Einsetzbarkeit	x	Laufwerk oder über einen Webbrowser (Smartphone oder Tablet mit HTML5-fähigen Browser)
5	Autorensystem	x	

Tabelle 4: Kriterienkatalog von Agorum Core Open

Wie bereits zuvor erwähnt, handelt es sich bei dieser Version des Systems um eine Open Source Lösung, die kostenlos ist. Regelmäßige Updates kosten allerdings einen kleinen Betrag.[108] Diese Möglichkeit wird als Zusatzfunktion im nächsten Punkt beschrieben.

Das zweite Kriterium der Priorität eins ist die Suchfunktion, die Agorum Core Open erfüllt. Folgende Möglichkeiten der Suche gibt es:

- Ähnlichkeitssuche: z.B. "Schmidt" suchen und auch "Schmitt" finden
- Wildcardsuche: z.B: "dokument*" sucht alle Worte, die mit "dokument" beginnen
- Volltextsuche: Suche innerhalb der Textinhalte von Dokumenten
- Attributsuche/Metadatensuche: Suche nach Attributen/Meta-Informationen der Objekte (kombinierbar mit der Volltextsuche)
- Numerische Suche: Suche von Zahlen
- Bereichssuche: von - bis, für Zahlenwerte aber auch für Datumsangaben
- Suche im Ablageort: Beschränkung auf diverse Ablageorte

Alle Such-Optionen sind miteinander kombinierbar und ein Text ist sofort nachdem er hochgeladen wurde, in die Suche integriert. Die Suchwörter werden aber nur in der Pro Version durch das System automatisch farbig hervorgehoben. Außerdem kann man eine Suche auch speichern, wenn man regelmäßig die neueste Version von bestimmten Dokumenten benötigt oder lediglich nach neuen Dateien zu einem bestimmten Thema suchen möchte.[109]

Das drittes Kriterium der ersten Priorität steltl die einfache Navigation dar. Diese ist durch einfache ergonomische Bedienung erfüllt, wie in Abbildung 6 zu erkennen ist. Man benötigt

[108] Vgl. Schulze, O. (2014)
[109] Vgl. agorum® Software GmbH (o.J.a)

z.B. nur einen Klick, um ein Dokument zu öffnen oder benötigt einen Rechts-Klick um ein Dokument bearbeiten zu können, etc.[110] Das Öffnen von Dokumenten durch nur einen Klick, erfüllt die als wichtig definierte Anforderung (vergleiche Kapitel 3), dass Dokumente nicht in einer komplizierten Ordnerstruktur abgelegt werden sollen. Eine komplexe Ordnerstruktur würde die Bedienung des Dokumentationssystems sehr stark im Hinblick auf Benutzer-freundlichkeit einschränken.

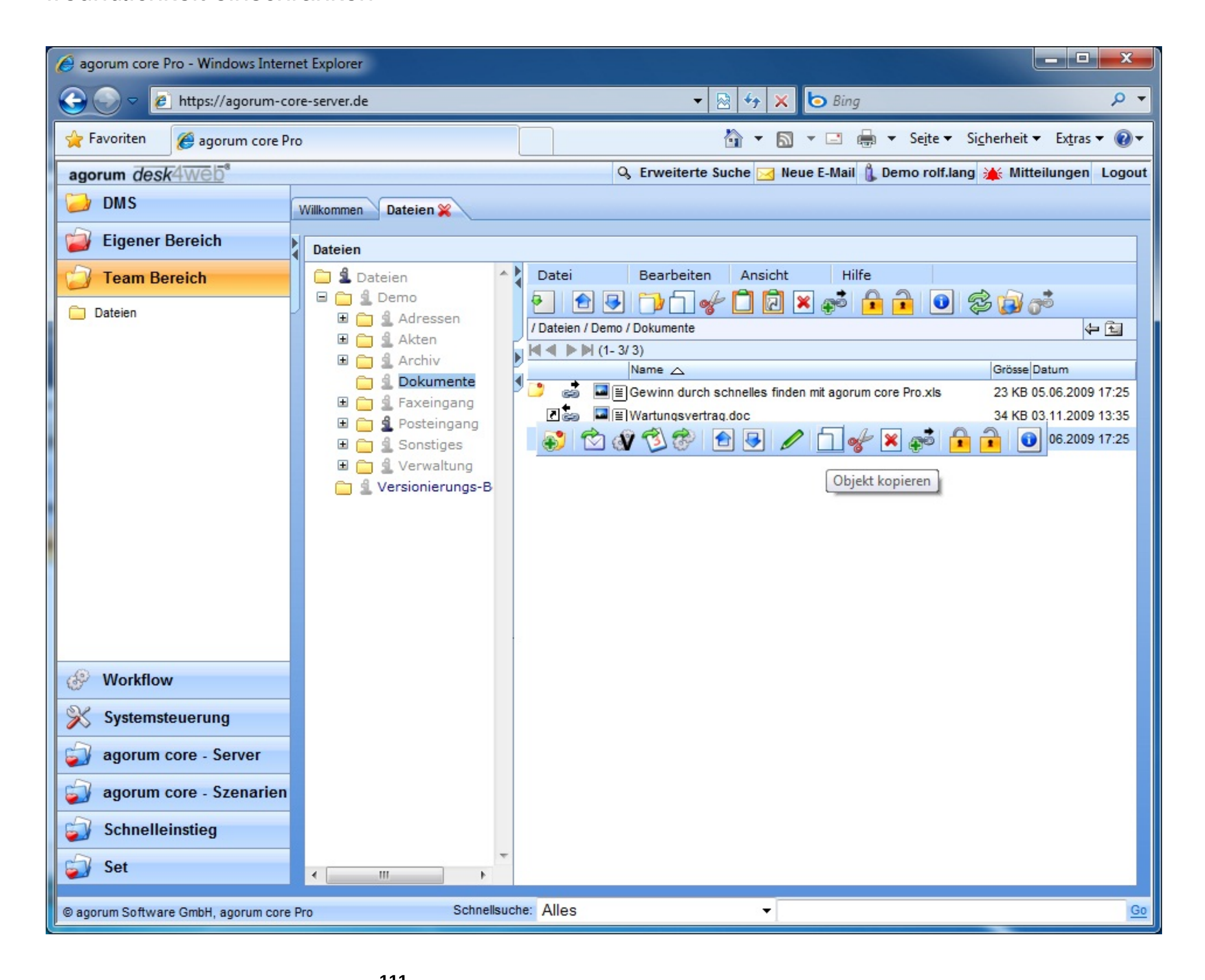


Abb. 6: Einfache Navigation[111]

Auf Agorum ist es möglich, Ordner, Dokumente, E-Mails, Wikis, Foren, Termine, Benutzer, Gruppen, und sogar eigene Objekttypen abzulegen. Alle gängigen Dokumenttypen werden unterstützt, wie zum Beispiel PDF, gif, jpg, jpeg,doc, docx, java, png, ppt, xls etc. Im Anhang 1 ist eine Liste mit allen unterstützen Dokument-Formaten zu finden.[112]

[110] Vgl. agorum® Software GmbH (o.J.d)
[111] Enthalten in: agorum® Software GmbH (o.J.d)
[112] Vgl. agorum® Software GmbH (o.J.c)

Somit sind alle Kriterien der ersten Priorität erfüllt. Als nächstes werden die Kriterien der zweiten Priorität untersucht.

Eine Einschränkung der Anzahl der User gibt es gibt. Es können also so viele User angelegt werden, wie benötigt.[113]

Die Funktionalität der Versionierung ist ebenfalls vorhanden. Wie in Abbildung 7 zu sehen ist, ist bei allen Dokumenten nachzuvollziehen, wer, wann, was geändert hat und auf alte Versionen zugreifen.[114]

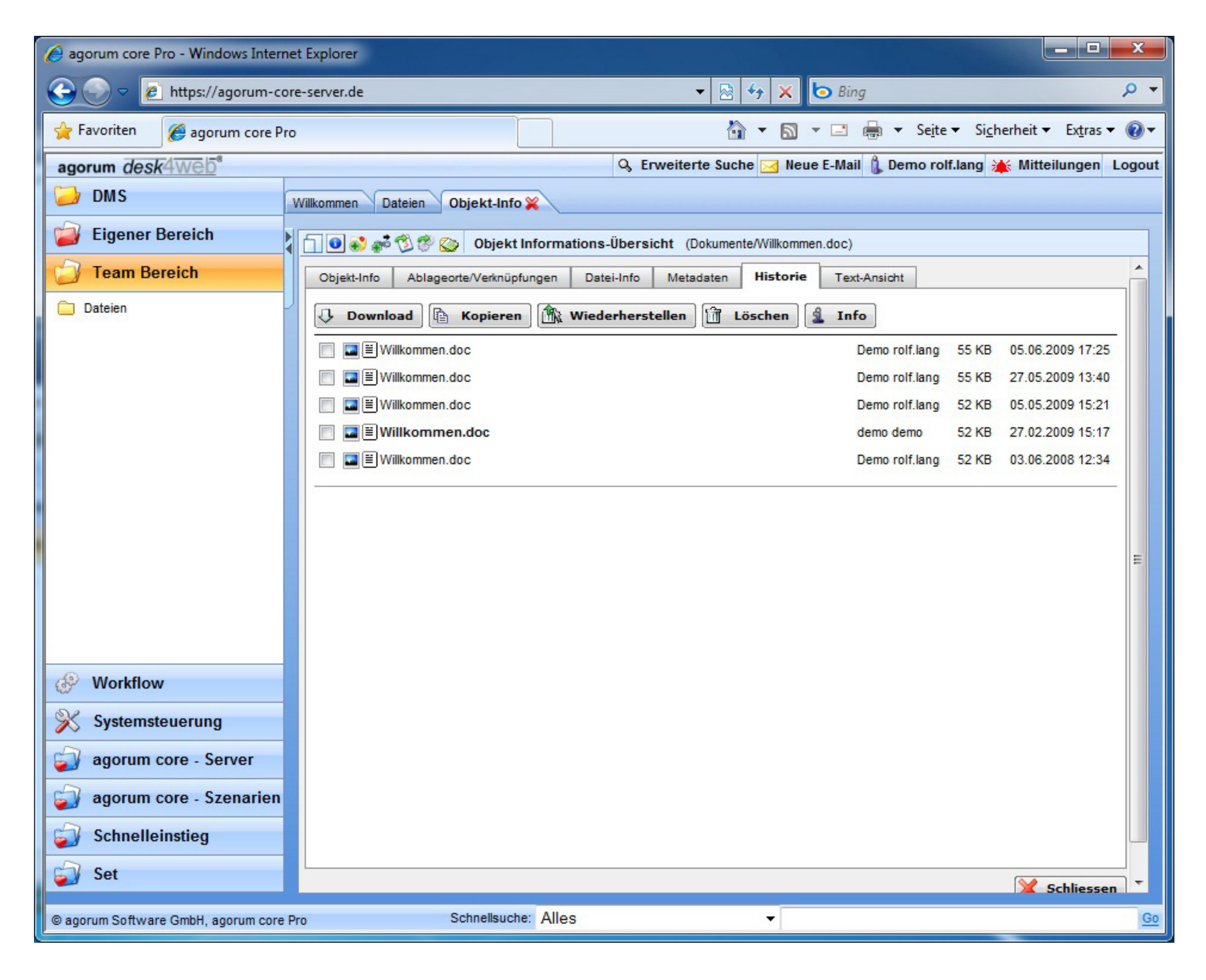


Abb. 7: Historie Agorum[115]

Die speicherbare Datenmenge ist nicht beschränkt. Es können so viele Dokumente wie gewünscht hochgeladen werden.[116]

[113] Vgl. Schulze, O. (2014)
[114] Vgl. agorum® Software GmbH (o.J.d)
[115] Enthalten in: Vgl. agorum® Software GmbH (o.J.d)
[116] Vgl. Schulze, O. (2014)

Somit sind alle Kriterien der Priorität zwei erfüllt.

Die Verbreitung beziehungsweise die Reife dieses Produkts bzw. des Unternehmens ist gut. Das Produkt wird zwar nur beinahe ausschließlich im deutschsprachigen Raum genutzt, hat aber bereits 10.000 Downloads. Das Produkt gibt es seit 2008, das Unternehmen bereits seit 1998.[117] Außerdem kann Agorum gute Referenzen aufweisen, wie z.B. Stadtverwaltungen oder Stadtwerke.[118]

Die Implementierung ist sehr einfach. Das System wird auf einem Server installiert und läuft dann auf Basis eines Netz-Laufwerks, welches im Windows Explorer angezeigt wird. Alle Funktionen, wie zum Beispiel Check-In/Check-Out können über den Explorer ausgeführt werden. Der Zugriff kann aber auch ohne spezielle Installation von einem beliebigen Arbeitsplatz über den Internet-Browser geschehen. Auch das führt zu einer einfachen Navigation.[119]

Auch bei der Datengröße pro Dokument gibt es bei Agorum keine Einschränkungen. Die Dokumente können beliebig groß sein.[120]

Alle Kriterien der Priorität drei sind somit erfüllt. Im nächsten Schritt werden alle Kriterien der Priorität vier untersucht.

Der Umgang mit Anhängen ist bei Agorum sehr gut. Es handelt sich nicht um ein Wiki mit Anhängen, sondern um ein Dokumentenverwaltungssystem. Diese Dokumente können aber wie Artikel in einem Wiki kommentiert oder verändert werden. Downloads für Templates etc. sind auf diesem Wege sicherlich einfacher. Ein Wiki kann allerdings auch angebunden werden.[121] Die Anhänge werden aber nicht wie bei einem Wiki direkt als Text angezeigt. Der User muss die Dokumente wie in einem Explorer öffnen. Wie die Speicherung der Dateien exakt erfolgt, konnte in Rahmen der Projektarbeit nicht herausgefunden werden. Es ist aber anzunehmen, dass die Speicherung in Form einer Datenbank auf dem Server, auf dem auch das Netz-Laufwerk läuft. Bei einem Aufruf über das Laufwerk im Explorer wird also direkt auf die Dateien zugegriffen, bei einem Aufruf über den Webbrowser greift dieser auf das Netz-Laufwerk zu.

Das Lizenzmodell ist GPL2.[122]

Ein Berechtigungssystem ist auch vorhanden und kann eingerichtet werden. Dies ist über sogenannte ACLs (AccessControlLists) möglich. Hier werden Benutzer zu bestimmten Listen

[117] Vgl. agorum® Software GmbH (o.J.e)
[118] Vgl. agorum® Software GmbH (o.J.f)
[119] Vgl. agorum® Software GmbH (o.J.b)
[120] Vgl. Schulze, O. (2014)
[121] Vgl. agorum® Software GmbH (o.J.b)
[122] Vgl. agorum® Software GmbH (o.J.a)

mit entsprechenden Zugriffsrechten hinzugefügt und verwaltet. Außerdem kann man Gruppen definieren. Zu den Berechtigungsstrukturen können auch Reports erstellt werden und es kann jederzeit eine Diagnose erstellt werden (Wer ist angemeldet, wer hat welches Recht, welche Dateien sind in Bearbeitung). Ein Dokument kann auch als unveränderbar deklariert werden, damit es niemand verändern kann.[123]

Somit sind alle Kriterien der Priorität eins bis vier erfüllt. Im nächsten Schritt wird die letzte Prioritätsgruppe (fünf) untersucht.

Es gibt es bei Agorum sehr viele Möglichkeiten das System selbst zu gestalten. Das System ist Ordner-basiert, d.h. die Struktur kann von dem User vollkommen selbst bestimmt werden. Zusätzlich ist es möglich, eigene Objekttypen mit eigenen Attributen zu bestimmen. Auch die Definition von Meta-Daten ist möglich, damit die Suche seiner eigenen Dokumente einfacher wird. Die Integration eigener Masken oder Designanpassungen zum Beispiel nach Berechtigung oder Gruppen sind auch möglich.[124]

Die Einsetzbarkeit ist vielfältig, da man über das Netz-Laufwerk am PC oder Laptop einfachen Zugriff im Explorer hat oder über einen Webbrowser, d.h. auch mit einem Smartphone oder Tablet mit HTML5-fähigen Browser auf das System zugreifen kann.[125]

Das Autorensystem ist ebenfalls vorhanden, die Autoren werden in der Historie angezeigt und über die Suchfunktion lässt sich auch nach Autoren suchen. Allerdings gibt es kein Profil, in dem alle Dokumente eines Autors angezeigt werden. [126] Dieses Kriterium ist also zum Teil erfüllt.

Somit erfüllt das System Agorum alle Kriterien ausreichend für die Benutzung in der Abteilung der Anwendungsentwicklung der .Versicherung . Im nächsten Punkt werden Zusatzfunktionen des Systems erläutert.

5.1.2 Zusatzfunktionen

Konsistenz:

Durch Links und Verknüpfungen von Objekten können gleiche Daten an unterschiedlichen Orten gleichzeitig sichtbar gemacht werden. Dadurch wird sichergestellt, dass Veränderun-

[123] Vgl. agorum® Software GmbH (o.J.b)
[124] Vgl. agorum® Software GmbH (o.J.b)
[125] Vgl. agorum® Software GmbH (o.J.b)
[126] Vgl. Schulze, O. (2014)

gen überall sichtbar sind und sich überall auswirken (zum Beispiel eine Verknüpfung zwischen zwei Excel-Tabellen), wodurch die Daten konsistent sind.

Sicherheit:

Der Zugriff auf das Portal erfolgt über eine SSL-Verschlüsselung und ist somit sicher verschlüsselt. Für gelöschte Objekte gibt es einen Papierkorb, in dem man alle gelöschten Dateien wieder herstellen kann. Die Backup-Funktion sorgt für regelmäßige Updates der gesamten Datenbank. Auch Berechtigungen und andere Einstellungen werden automatisch gesichert.

Prozesse:

Es gibt die Möglichkeit, bestimmte Dateien mit Prozessen zu verknüpfen. Zum Beispiel lassen sich Dokumente auf Termine legen oder es können automatische Mitteilungen bei Änderungen von Dokumenten versendet werden (z.B. mit Outlook). Eine weitere mögliche Automatisierung stellt die automatische Archivierung nach x Tagen dar.

Integration:

Ein CRM System kann in Agorum integriert werden, sodass Kunden (mit Adressen etc) mit bestimmten Akten oder Dokumenten verknüpft werden können. Es gibt auch die Möglichkeit Emails zu bündeln und Konversationen zu bestimmten Akten zu archivieren. Auch Dritt-Programme wie Converter, Fax- oder Scan-Programmen können in Agorum eingebunden werden. Über SOAP-Webservices oder das Laufwerk kann Agorum auch in interne, eigene Programme integriert werden.

Reports:

In Agorum gibt es einen integrierten Reportgenerator, der verschiedenste Reports (Berechtigungsstruktur, Nutzungshäufigkeit etc.) erstellen und exportieren kann.[127]

Die eben beschriebenen Funktionen sind alle in der Open-Source Version enthalten und sind kostenlos. Es gibt auch eine core pro Version mit Zusatzmodulen wie Workflow, E-Mail Archiv etc. Alle wesentlichen Unterschiede sind in einer Broschüre auf der Webpage von Agorum zu finden.[128]

Außerdem gibt es ein Abonnement für regelmäßige Updates der Versionen. Der Preis hierfür ist abhängig von der Anzahl der Nutzer. Das Update Abonnement kostet für 150 Benutzer

[127] Vgl. agorum® Software GmbH (o.J.b)
[128] Vgl. agorum® Software GmbH (2014)

Jährlich 1.690,00 EUR + MwSt. Ein Einmal-Update für 150 Benutzer kostet 1.014,00 EUR + MwSt.[129]

Agorum erfüllt alle Kriterien und ist sehr für die Verwendung der .Versicherung Versicherung zu empfehlen.

5.2 Alfresco

Alfresco ist eine Open Source Plattform, welche für das Managen der unterschiedlichen Dokumente von Unternehmen vorgesehen ist. Es ist hier zu beachten, dass Alfresco unterschiedliche Tools anbietet. Zum einen bieten sie ein On-Premise Geschäftsmodell an, welches die Unternehmensdaten auf dem Server der Unternehmen speichert. Des Weiteren bieten sie eine Cloud Version an. Hier werden die Unternehmensdaten auf dem AWS-Server in Frankfurt gespeichert.[130] Auch gibt es ein kostenloses Angebot, auf das später eingegangen wird. [131]

5.2.1 Kriterien

Priorität	Kriterium	Alfresco	Kommentare
1	Open Source	x	
1	Suchfunktion	x	Volltextsuche
1	Einfache Navigation	x	
1	Ablage von Dateien	x	
2	70-150 User	größer	mehrere User möglich
2	Versionierung	x	
2	Speicher Datenmengen <20.000	größer	größere Datenmengen sind möglich
3	Verbreitung/Reife	x	
3	einfache Implementierung	x	Webanwendung, einmalige Installation genügt
3	Datengröße	x	unbegrenzt
4	Handling von Anhängen	/	
4	Lizenzmodell	x	genaue Preisauskunft vom Hersteller nicht möglich
4	Berechtigungssystem	x	
5	Gestaltungsmöglichkeiten	x	

[129] Vgl. Schulze, O. (2014)
[130] Vgl. Pauka J. (2015)
[131] Vgl. Alfresco Software, Inc. (2015a)

5	Einsetzbarkeit	x	
5	Autorensystem	x	

Tabelle 5: Kriterienkatalog von Alfresco

Wie bereits erwähnt, ist Alfresco ein Open Source Produkt. Das heißt jedoch nicht, dass es kostenlos zu verwenden ist. Ein Mitarbeiter von Alfresco beschreibt ihr Produkt als eine kommerzielle Open-Source Lösung. Detaillierte Preisauskunft ist nicht möglich, da die Preisinformation nicht offiziell ist. Eine Preisanfrage bei Alfresco muss mit einem gezielten Projekt verbunden sein, erklärt Janusz Pauka, Mitarbeiter bei Alfresco. Wichtig ist jedoch die Tatsache, dass es Open Source ist und somit ein frei verfügbarer Open Source Code verwendet werden kann. Somit ist Das Kriterium Open Source mit Wichtigkeit eins erfüllt.[132]

Eine weitere notwendige Funktion mit Priorität eins ist die Suchfunktion. Auch diese ist bei Alfresco aufzufinden. Sie arbeiten mit einer Volltextsuche namens Apache Lucene. [133] Auch Apache Lucene ist ein Open Source Projekt, welches auf nahezu jeder Applikation möglich ist. [134]

Das Tool zur Betreibung von Wissensmanagement besitzt eine gewisse Benutzerfreundlichkeit. Aufgrund einer ähnlichen Funktion wie bei einem virtuellen externen Netzlaufwerk ist es verständlich für die Benutzer. Dies führt zu einem kürzeren Schulungsaufwand, weniger Kosten und einer schnellen Einführung des Systems. Außerdem bietet Alfresco einen Kundensupport, welcher weltweit zur Verfügung steht. [135] Des Weiteren ist ein einfaches Hochladen der Dokumente über Drag and Drop möglich. Somit ist das Kriterium „einfache Navigation" mit Priorität eins durch Alfresco erfüllt.

Ein weiterer wichtiger Punkt ist das Hochladen verschiedener Dateien. Auch hier ist bei Alfresco eine Vielzahl an Dateien möglich. Neben Standarddokumenten, wie docs, xls, jpg, zip oder auch pdfs sind deutlich mehr möglich. Somit ist auch das letzte Kriterium mit Priorität eins erfüllt.

Bei der Anzahl der User des Tools gibt es mehrere Differenzierungen. Unter dem Modell On-Premise gibt es zwei verschiedene Angebote. *Alfresco One Departmental* kann mit bis zu 300 Usern genutzt und blockweise um 25 Benutzer erweitert werden. Alle diese Anwender haben einen Zugang zu den Dokumenten. 100 Benutzer der 300 haben die Möglichkeit die Dokumente auf einem Amazon Server zusätzlich abzuspeichern bzw. synchronisieren zu

[132] Vgl. Pauka J. (2015)
[133] Vgl. Bodgan M. (o.J.)
[134] Vgl. Apache Software Foundation (o.J.)
[135] Vgl. Alfresco Software, Inc. (2015b)

lassen - ein sogenanntes Hybridmodell. [136] (Ein Hybridmodell ist ein CAD-Modell, welches aus Daten unterschiedlicher Herkunft zusammengesetzt wird.[137]) Die zweite Version des On-Premise Angebots ist das *Alfresco One Enterprise Modell*. Hier können bis zu 1000 User mit einer blockweisen Erweiterung von 100 Benutzern integriert werden.

Bei der Cloud Version von Alfresco kommt nur ein Modell für die :vERSICHERUNG in Frage, da das andere zu wenige User umfasst. Das *Enterprise Network* kann 500 User mit einer Erweiterung der Anwenderanzahl von einem User aufnehmen. [138]

Des Weiteren ist eine Versionierung mit Priorität zwei behaftet. Bei Alfresco werden bei jeder Änderung der Dokumente automatisch neue Versionen erzeugt und gleichzeitig Kopien der vorherigen Versionen gespeichert. Außerdem gibt es eine Angabe, welcher User die Änderungen umgesetzt hat.[139] Hier wird deutlich, dass eine Nachverfolgung der verschiedenen Versionen möglich ist. Die Tatsache, dass es eine Angabe gibt, wer welche Änderungen gemacht hat, deckt auch sofort das Kriterium des Autorensystems mit Priorität fünf ab.

Der Speicherplatz bei *Alfresco One Premise* ist unbegrenzt, da die Dateien auf den Servern des Unternehmens abgelegt sind. Nachdem es hier keine begrenzte Speichermenge gibt, ist dieses Kriterium mit Priorität zwei erfüllt. [140]

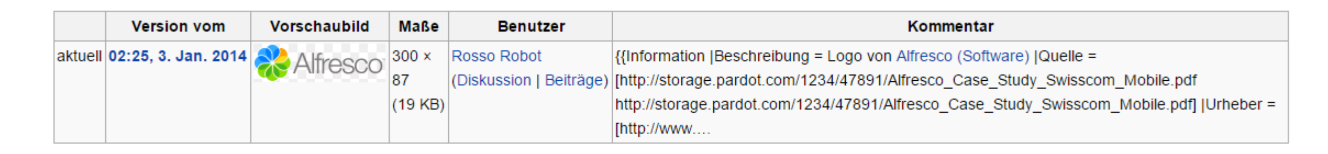
Alfresco ist mittlerweile stark verbreitet. Auf der Firmenhomepage sind Kunden von NASA bis zu der Fluggesellschaft KLM vertreten. Insgesamt hat Alfresco bis zu 11 Millionen User weltweit vorzuweisen. [141] Dies ist ein starkes Merkmal für die Verbreitung von Alfresco und somit ist dieses Kriterium mit Priorität drei erfüllt.

Dieses Tool ist eine Webanbindung. Das bedeutet die Installation von Alfresco genügt einmalig und jeder User kann via eines Webbrowsers darauf zugreifen. Dies ist eine deutliche Erleichterung, nicht für jeden User ein einzelnes Tool implementieren zu müssen. Das bedeutet, dass die einfache Implementierung für die User gegeben ist und das Kriterium mit Priorität drei auch erfüllt ist.

Aus demselben Grund wie die speicherbare Datenmenge ist auch die Datengröße der einzelnen hochgeladenen Dokumente unbegrenzt. Somit ist die letzte der Priorität drei erfüllt, da dies für die User deutliche Erleichterung darstellt.

[136] Vgl. Werner B. (2009)
[137] Vgl. Blien R. (2009)
[138] Vgl. Pauka J. (2015)
[139] Vgl. Alfresco Software, Inc. (2015b)
[140] Vgl. Pauka J. (2015)
[141] Vgl. Alfresco Software, Inc. (2015c)

Der Umgang mit Anhängen ist in dem vorliegenden Kriterienkatalog mit Priorität vier versehen. Es ist möglich den Speicherort über Einstellungen in der Datei ändern. Der User kann also selbst entscheiden wo er seine hochgeladenen Dateien speichern möchte. Abbildung 8 zeigt die Dateiversion von gespeicherten Daten bei Alfresco. Es ist somit nicht möglich die Dateien als Vollbild zu sehen und dieses Kriterium kann bei Alfresco nicht erfüllt werden.



	Version vom	Vorschaubild	Maße	Benutzer	Kommentar
aktuell	02:25, 3. Jan. 2014	Alfresco	300 × 87 (19 KB)	Rosso Robot (Diskussion | Beiträge)	{{Information |Beschreibung = Logo von Alfresco (Software) |Quelle = [http://storage.pardot.com/1234/47891/Alfresco_Case_Study_Swisscom_Mobile.pdf http://storage.pardot.com/1234/47891/Alfresco_Case_Study_Swisscom_Mobile.pdf] |Urheber = [http://www.…

Abb. 8: Dateiversion bei Alfresco [142]

Die Updates bei Alfresco werden per Subskription für zwölf Monate verkauft. Wie bereits erwähnt, kann man die einzelnen Modelle käuflich erwerben. Eine genaue Preisauskunft ist jedoch vom Hersteller nicht möglich. [143] Die Nutzer können selbst entscheiden, welches Support Paket sie wählen wollen siehe Abbildung 9. Es ist auch eine kostenlose Version von Alfresco vorhanden (Community Edition). Diese bekommt jedoch in keiner Weise Unterstützung von Alfresco und hat auch weniger Funktionen anzubieten. [144] Dem zufolge ist ein weiteres Kriterium der Priorität vier erfüllt, da Lizenzen zur Verfügung stehen.

[142] Mit Änderungen entnommen aus: Wikipedia (2014)
[143] Vgl. Pauka J. (2015)
[144] Vgl. Alfresco Software, Inc. (2015a)

Vergleichsmatrix

	Community	Departmental	Enterprise
Autorisierte Supportkontakte	0	2	3
Dokumentation	✓	✓	✓
Foren	✓	✓	✓
Zertifizierte Binärdateien		✓	✓
Kritische Meldungen		✓	✓
Knowledge Base		✓	✓
Zugriff auf Support per Telefon/Web		✓	✓
Service-Packs		✓	✓
Support zu normalen Geschäftszeiten		✓	✓
Rund-um-die-Uhr-Fehlerbehebung bei Schweregrad 1			✓

Abb. 9: Vergleichsmatrix des unterschiedlichen Supports der einzelnen Pakete[145]

Durch die unten aufgelisteten Benutzerrechte wird auch das letzte Kriterium der Wichtigkeit vier erfüllt:

Administrator:

Ein Administrator hat vollen Zugriff auf die Inhalte

Site Manager:

Sie haben vollen Zugriff auf alle Inhalte der eigenen Seite

Konsument:

Die Consumer dürfen die Inhalte nur lesen.

Editor:

Diese dürfen den Inhalt anderer User bearbeiten, jedoch keinen eigenen erstellen.

Beitragender:

Sie dürfen Inhalte erstellen, jedoch keine anderen Inhalte löschen.

Mitarbeiter:

Die Mitarbeiter dürfen die eigenen Inhalte und die anderer bearbeiten.

Koordinator:

Koordinatoren haben dieselben Rechte wie die Administratoren. [146]

[145] Mit Änderungen entnommen aus: Alfresco Software, Inc. (2015d)

[146] Vgl. Unitversität Kassel - IT Servicezentrum (o.J.)

Auch Gestaltungsmöglichkeiten sind bei Alfresco gegeben. Es gibt einen Satz von Design-Vorlagen, welche sich die Benutzer des Tools aussuchen können. Es können maßgeschneiderte, dynamische Web-Anwendungen entwickelt werden ohne ganz am Anfang starten zu müssen.[147] Somit ist auch dieses Kriterium mit Priorität fünf erfüllt.

Aufgrund der Tatsache, dass Alfresco über einen Webbrowser läuft und die User nur einen Internetzugang benötigen, um auf dieses Tool zugreifen zu können, ist es neben einem normalen PC auch auf einem mobilen Endgerät nutzbar. Also ist das Kriterium der Einsetzbarkeit auch gegeben.

Das letzte Kriterium, welches mit Priorität fünf behaftet ist, ist das Autorensystem. Auch dieses Kriterium erfüllt Alfresco. Wie in Abbildung 8 bereits sichtbar geworden ist, wird bei Alfresco der User genau angezeigt. Außerdem ist es möglich weitere Beiträge der einzelnen Nutzer nachzuverfolgen.

5.2.2 Zusatzfunktionen

Alfresco hat neben den Kriterien des Kriterienkatalogs auch noch Zusatzfunktionen zu bieten. Eine dieser Funktionen ist die Interaktion mit anderen sozialen Netzwerken. So können Alfresco Benutzer mit „Gefällt mir" markieren und anderen Usern „folgen". Auch können Nutzer ihre Daten über andere Kanäle, wie beispielsweise Facebook, Youtube, Twitter oder auch Flickr veröffentlichen. Durch diese Funktion wird das Tool persönlicher. Außerdem kann man mittlerweile Standardanwendungen, wie Microsoft Office, Google Text & Tabellen, Apple iWork usw. mit einbinden. Die letzte Zusatzfunktion von Alfresco ist eine neue Anwendung in der Cloud. Hier können die User auch beispielsweise mit Kunden an den Dokumenten arbeiten.[148]

Alles in allem lässt sich sagen, dass Alfresco ein interessantes Tool ist, welches mittlerweile bei vielen Unternehmen großen Anklang findet. Die Tatsache, dass Alfresco nicht kostenlos ist (außer die Community Edition) bringt nicht nur Nachteile mit sich. Dadurch haben die Unternehmen und deren Nutzer keinerlei Probleme bezüglich des Supports. Zusammenfassend lässt sich sagen, dass Alfresco empfehlenswert ist, sobald ein Unternehmen bereit ist Geld in ein Wissensmanagementtool zu investieren.

[147] Vgl. Alfresco Software, Inc. (2015d)
[148] Vgl. Alfresco Software, Inc. (o.J.)

5.3 DokuWiki

DokuWiki ist eine Plattform zum Teilen und Bearbeiten von Dateien, sowie zur Dokumentation, in Form eines Wikis. Die Software wurde im Jahr 2004 von dem deutschen Programmierer Andreas Gohr entwickelt.[149] DokuWiki ist unter der GNU GPL Lizenz Version 2 lizensiert.[150] Die Software wird regelmäßig aktualisiert und neue Releases werden herausgebracht.[151]

5.3.1 Kriterien

Im Folgenden wird die Wiki Software DokuWiki auf die Anforderungen des hier dargestellten Kriterienkataloges hin analysiert. Anschließend werden noch weitere Funktionen, die das Tool bietet, aber die nicht im Kriterienkatalog vorkommen, analysiert.

Priorität	Kriterium	DokuWiki	Kommentar
1	Open Source	X	Kostenlos und Quellcode verfügbar
1	Suchfunktion	X	Index-basierte Volltextsuche
1	Einfache Navigation	X	Breadcrumbs-Navigation
1	Ablage von Dateien	X	Anhänge (Bilder, Videos, pdfs, …)
2	70-150 User	X	keine Begrenzung
2	Versionierung	X	Ältere Versionen werden aufbewahrt
2	Speicher Datenmengen <20.000	X	keine Begrenzung
3	Verbreitung/Reife	X	Sehr verbreitet, regelmäßige Releases
3	einfache Implementierung	X	Anleitungen verfügbar

[149] Vgl. DokuWiki (2012a)
[150] Vgl. DokuWiki (2014a)
[151] Vgl. DokuWiki (2014f)

3	Datengröße	X	keine Begrenzung
4	Handling von Anhängen	X	Durchsuchen von Anhängen
4	Lizenzmodell	X	GNU GPL v2
4	Berechtigungssystem	X	Zugriffskontrolle durch Listen
5	Gestaltungsmöglichkeiten	X	Vielfältig – Templates, Plugins
5	Einsetzbarkeit	?	Webanwendung
5	Autorensystem	?	Autoren von Änderungen sichtbar/ Anhänge?

Tabelle 6: Kriterienkatalog von DokuWiki

Da das Wiki als Open Source Software verfügbar ist, kann es kostenlos runtergeladen werden. Für jegliche Nutzung oder Anpassung des Sourcecodes muss keinerlei Gebühr bezahlt werden. Der Sourcecode ist für jeden auf der Webseite von DokuWiki zugänglich und kann dort heruntergeladen werden. Für Unternehmen gibt es auf freiwilliger Basis die Möglichkeit einen beliebigen Betrag an die Entwickler zu spenden, da diese viel Zeit für Entwicklung und Wartung des Tools aufbringen.[152]

Die Wiki Software DokuWiki gehört laut Chip.de zu den drei besten Anbietern von Wikis.[153] Auch eine Statistik der Wiki-Vergleichs-Seite wikimatrix.org zeigt, dass DokuWiki in den letzten 30 Tagen mit Abstand am häufigsten angeklickt wurde (1. DokuWiki 3255 Klicks, 2. TWiki 2211 Klicks), was für eine große Bekanntheit des Wikis spricht. Die nachstehende Abbildung zeigt den Vergleich.[154]

[152] Vgl. DokuWiki (2014e)
[153] Vgl. Peters, M.(2014)
[154] Vgl. CosmoCode GmbH (2014)

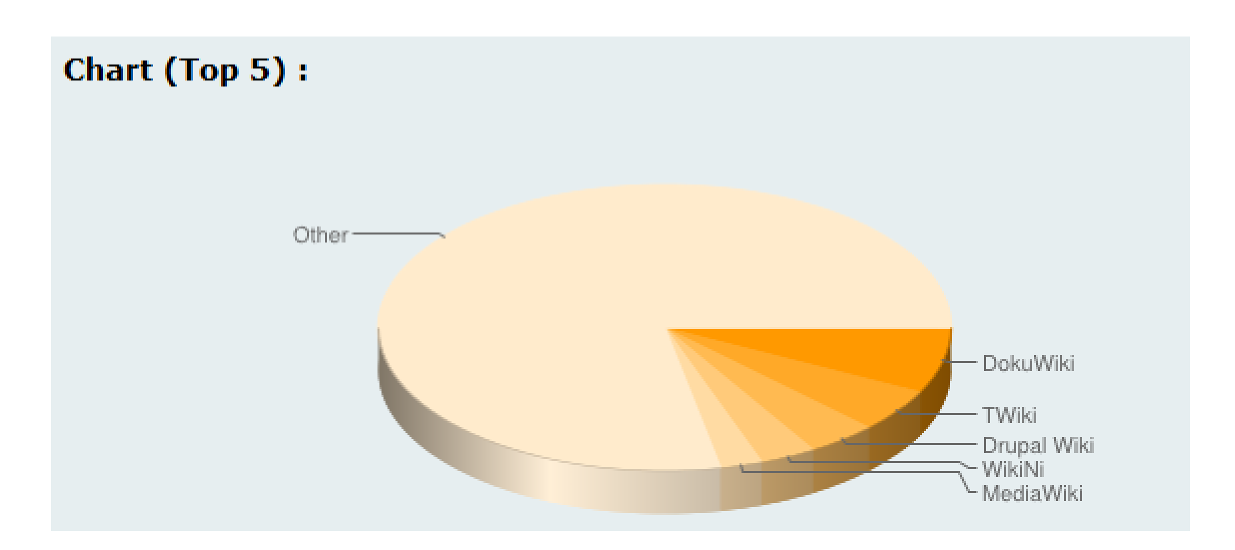


Abb. 10: Vergleich Wiki Software auf Basis registrierter Klicks innerhalb eines Monats[155]

Das Wiki kann in verschiedene Namensräume unterteilt werden. Dadurch entsteht eine Struktur die mit einer Verzeichnis- oder einer Ordnerstruktur vergleichbar ist.[156] Um sich in dem Wiki zurechtzufinden, bietet DokuWiki als Suchfunktion eine Indexbasierte Volltext-Suche.[157] Durch die Verwendung eines Indexes ist es möglich, Datensätze anhand von Suchbegriffen zu sortieren, wodurch diese Art der Suche wesentlich schneller ist als beispielsweise eine sequentielle Suche.[158] Bei Nutzung der Standardvorlage des Wikis befindet sich in der Kopfzeile ein Suchfeld. Die Suche berücksichtigt keine Groß und Kleinschreibung und es gibt verschiedene Möglichkeiten die Suchergebnisse einzugrenzen. So bietet die Software zum Beispiel die Möglichkeit bestimmte Begriffe in der Suche auszuschließen oder nur nach Fragmenten zu suchen.

Um eine einfache Navigation zu gewährleisten nutzt die Software die Breadcrumbs Navigation (deutsch: Brotkrümel-Navigation). Durch diese Art der Navigation wird die Nutzung des Tools in Form eines linearen Pfads dargestellt.[159] DokuWiki bietet zwei Arten der Breadcrumbs Navigation:[160]

1. Tracking Breadcrumbs

2. Hierarchical Breadcrumbs

[155] Enthalten in: CosmoCode GmbH (2014)
[156] Vgl. DokuWiki (2015c)
[157] Vgl. DokuWiki (2015a)
[158] Vgl. Marunde, G. (2003), S. 49
[159] Vgl. Arndt, Henrik (2006), S. 237
[160] Vgl. DokuWiki (2013b)

Tracking Breadcrumbs bedeutet, dass eine Liste der Seiten angezeigt wird, die vor kurzem besucht wurden, wobei die Anzahl der angezeigten Seiten in den Optionen eingestellt werden können. Diese Funktion ist standardmäßig eingestellt und für Seiten mit einer flachen Namensraumhierarchie besonders sinnvoll.[161] Die Funktion der Hierarchical Breadcrumbs hingegen muss bei Bedarf erst aktiviert werden und eignet sich für Wikis mit einer tiefen Seitenstruktur, denn hier wird dem Nutzer angezeigt, auf welcher Ebene er sich gerade in der organisatorischen Hierarchie der Seite befindet.[162]

[161] Vgl. DokuWiki (2013b)
[162] Vgl. Herrington, J. D./ Lang, J. W. (2006), S.23

Das Wiki DokuWiki ermöglicht das Hochladen von verschiedenen Dateien wie Bildern, Videos oder Dokumenten. Hochgeladene Bilder und Flash Dateien werden direkt auf der Seite angezeigt. Andere Anhänge werden allerdings nicht zu einer bestimmten Seite hinzugefügt sondern in einem zentralen Speicher abgelegt und nur auf der Seite verlinkt.[163] Die hochgeladenen Medien oder Dokumente können den verschiedenen Namensräumen zugeordnet werden. Der Zugriff auf die Anhänge kann durch eine Zugriffsrechtevergabe geregelt werden. Für hochgeladene Bilder gibt es die Möglichkeiten, diese direkt in eine Seite einzugliedern und die Größe des Bildes automatisch anpassen zu lassen.[164] Der nachstehende Screenshot (Abbildung 11) zeigt den Media Manager. Hier können Anhänge hochgeladen und gespeichert werden. Außerdem hat der User hier die Möglichkeit. die Anhänge zu durchsuchen.

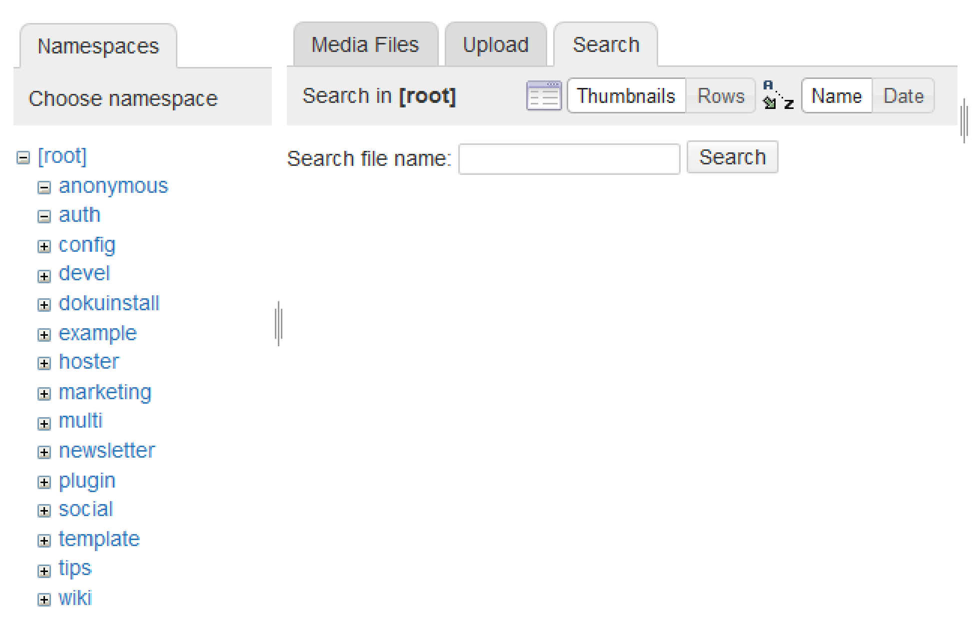


Abb. 11: Screenshot des Media Managers im DokuWiki[165]

Für die Anzahl der User des Wikis gibt es keine Begrenzung.[166] Grundsätzlich kann auch jeder Benutzer alles einsehen und bearbeiten. Erst wenn die Zugriffskontrolle des Wikis aktiviert ist, besteht die Möglichkeit sich für die Nutzung des Wikis zu registrieren und einzuloggen. Das Wiki bietet allerding auch die Funktion, dass nur Administratoren neue Nutzer hinzufügen können, sodass das Wiki nicht von jedem einsehbar ist und firmenintern

[163] Vgl. DokuWiki (2014h)
[164] Vgl. DokuWiki (o. J.)
[165] Mit Änderungen entnommen aus: DokuWiki (2015b)
[166] Vgl. Henke A. (2015)

bleibt.[167] Es gibt insgesamt fünf verschiedene Zugriffsreche für die Nutzer die in ihrer Wertigkeit aufsteigend sortiert sind: lesen, editieren, anlegen, hochladen, löschen. Das höhere Zugriffsrecht enthält hierbei jeweils die darunterliegenden. Die Zugriffsrechte beziehen sich entweder auf die Seiten oder auf die Namensräume.[168]

Eine Versionierung in Form der Aufbewahrung von älteren Versionen ist im DokuWiki ebenfalls möglich. Sobald der Inhalt einer Seite verändert wird, wird eine neue Version erstellt. Für ältere Versionen gibt es ein Verzeichnis (Attic-Verzeichnis), in welches diese verschoben und dort aufbewahrt werden. Dieses Verzeichnis wird durch den Button „Ältere Versionen" angezeigt. Ältere Versionen werden nicht automatisch gelöscht, sondern müssen manuell gelöscht werden.[169] Das Wiki bietet auch eine Funktion (diff Funktion), die verschiedene Versionen vergleicht und die Änderungen der Versionen darstellt.[170] Eine weitere Funktion die die Änderungen darstellt ist die Funktion „Letzte Änderungen" (in der untenstehenden Abbildung dargestellt). Diese Funktion hingegen zeigt eine Liste mit den letzten Änderungen aller Seiten im Wiki an. Hierbei wird allerdings nur die allerletzte Änderung berücksichtigt. Bei dieser Funktion sind neben Bearbeitungszeitpunkt auch der bearbeitende Nutzer und die Veränderungen auf der Seite ersichtlich. Die letzten Änderungen können entweder für einen bestimmten Namensraum oder für das gesamte Wiki angezeigt werden.[171]

Abb. 12: Letzte Änderungen[172]

Für den Umfang des Wikis (Anzahl der Seiten, hochgeladene Bilder oder Dokumente) gibt es von der Software keine Begrenzungen. Auch die Größe der hochzuladenden Anhänge ist

[167] Vgl. DokuWiki (2013d)
[168] Vgl. DokuWiki (2013f)
[169] Vgl. DokuWiki (2011a)
[170] Vgl. DokuWiki (2011b)
[171] Vgl. DokuWiki (2013c)
[172] Mit Änderungen entnommen aus: DokuWiki (2014d)

nicht begrenzt. Die einzige Begrenzung hierbei ist das eigene Dateisystem und dessen Speicherplatz.[173]

Die Installation der Wiki Software wird als sehr einfach beschrieben und ist auf der Webseite von DokuWiki in sechs Schritten erklärt. Die Seite stellt außerdem ein Video zur Verfügung, in dem die Installation schrittweise gezeigt wird. Um sich besser mit dem Programm zurechtzufinden, stellt DokuWiki auf seiner Webseite ein Handbuch zur Verfügung, in dem Grundlegendes aber auch viele Details beschrieben und erklärt werden.[174] Für weitere Fragen gibt es auf der Webseite von DokuWiki ein Forum, in dem die Administratoren schnelle Antworten zu jeglichen Fragen und Problemen liefern. Voraussetzungen für die Installation eines MediaWikis sind zum einen ein Webserver, der PHP unterstützt und das verwendete PHP muss mindestens Version 5.1.2 oder höher sein. Außerdem wird ein aktueller Web-Browser benötigt.[175] Eine Besonderheit des DokuWikis ist, dass keine Datenbank für dessen Nutzung benötigt wird, da DokuWiki reine Textdateien (.txt) verwendet.[176]

Die Gestaltungsmöglichkeiten eines DokuWikis sind vielfältig durch Templates und Plugins vorhanden. Neben der Möglichkeit sein eigenes Template zu erstellen, gibt es insgesamt 125 Templates, die auf der Seite von DokuWiki heruntergeladen werden können. Diese Templates ermöglichen beispielsweise die Gestaltung des Farbschemas oder des Aufbaues der Menüleisten.[177] Mit Plugins können die Funktionen des Wikis erweitert werden. Auch hier gibt es neben der Möglichkeit ein eigenes Plugin zu schreiben 1087 verfügbare Plugins, mit denen das Wiki erweitert werden kann.[178] Zudem lassen sich in einem DokuWiki zur Zeit über 30 Sprachen einstellen.[179]

Zu dem Autorensystem lässt sich sagen, dass es möglich ist zu sehen, welche Person eine Seite erstellt hat und welche Personen Änderungen auf einer Seite getätigt hat. Ob dargestellt wird, wer welchen Anhang hochgeladen hat ist auf der Testseite von DokuWiki nicht ersichtlich geworden. Auch eine weitere Einsatzmöglichkeit des Wikis als mit dem Webbrowser wurde bei der Recherche nicht ersichtlich.

[173] Vgl. Henke A. (2015)
[174] Vgl. DokuWiki (2014b)
[175] Vgl. DokuWiki (2012b)
[176] Vgl. DokuWiki (2015a)
[177] Vgl. DokuWiki (2014c)
[178] Vgl. DokuWiki (2013e)
[179] Vgl. DokuWiki (2015a)

5.3.2 Zusatzfunktionen

Neben den geforderten Funktionen aus dem Kriterienkatalog, gibt es noch weitere Funktionen in einem DokuWiki. So besteht zum Beispiel die Möglichkeit durch sogenannte InterWiki Links Verknüpfungen zu anderen Wikis in das eigene Wiki mit einzubringen.[180]

Um die Nutzerfreundlichkeit zu erhöhen gibt es im DokuWiki die Funktion des „Section Editing". Durch diese Funktion wird dem Nutzer ermöglicht, lediglich einen bestimmten Teil beziehungsweise einen Absatz einer Seite zu bearbeiten, was besonders bei sehr großen und umfangreichen Seiten die Bearbeitung erleichtert. Außerdem wird eine Seite, die gerade bearbeitet wird für andere Nutzer zur Bearbeitung gesperrt. Dadurch werden Speicherkonflikte vermieden. Des Weiteren offeriert die Wiki-Software eine optionale Überprüfung der Rechtschreibung sowie die automatische Erstellung eines Inhaltsverzeichnisses.[181] Um die Bearbeitung des Wikis, sowie die Navigation innerhalb des Wikis einfach zu gestalten, gibt es jede Menge Tastaturkürzel. Diese Tastaturkürzel ermöglichen zum einen das Öffnen von verschiendenen Seiten in einem anderen Modus (z. B. Lesemodus oder Bearbeitungsmodus) aber auch das Bearbeiten des Textes (z. B. kursive, fette, unterstrichene Wörter). Die Anwendung der Kürzel kann sich allerdings bei unterschiedlichen Browsern unterscheiden.[182] Für die Bearbeitung von Texten ohne Tastaturkürzel gibt es eine Formatierungs-Knopfleiste, die auf der Formatierungs-Knopfleiste von MediaWiki basiert und genauso wie die Tastaturkürzel gängige Buchstaben beziehungsweise Symbole verwendet (z.B. B für fett, I für kursiv, U für unterstrichen).[183]

Zusammenfassend lässt sich sagen, dass die Wiki-Software DokuWiki den Anforderungen des Kriterienkalatoges im Großen und Ganzen entspricht und somit von diesem Standpunkt aus betrachtet eine geeignete Software darstellen würde.

Außerdem bietet die Wiki-Software DokuWiki neben den bisher genannten Funktionen noch einige weitere Funktionen. Zum einen hat der Nutzer dadurch sehr viele Möglichkeiten sich sein Wiki nach Wunsch zu gestalten, aber auf der anderen Seite kann es einige Zeit in Anspruch nehmen sich mit den vielen Funktionen auseinanderzusetzen und die wichtigen beziehungsweise notwendigen Funktionen herauszufiltern. Die Analyse aller Funktionen von DokuWiki würde den Rahmen dieser Arbeit allerdings überschreiten und die genauen Funktionalitäten aller Funktionen sind auf der Internetseite von DokuWiki nicht ganz ersichtlich.

[180] Vgl. DokuWiki (2015a)
[181] Vgl. DokuWiki (2015a)
[182] Vgl. DokuWiki (2013a)
[183] Vgl. DokuWiki (2014g)

5.4 Media Wiki

Media Wiki ist eine freie Serverbasierte Wiki Software. Eine Wiki Software ist ein Hypertextsystem für Webseiten, in dem die Inhalte von den Benutzern nicht nur gelesen, sondern auch geändert werden können. Das Ziel eines solchen Wikis ist, die Erfahrung und das Wissen der verschiedenen Nutzer gemeinsam zu sammeln (kollektive Intelligenz) und dementsprechend verständlich für alle zu dokumentieren. Die wohl bekannteste Plattform, welche auf Media Wiki basierend ist, ist Wikipedia. [184]

[184] Vgl. Media Wiki (2014e)

5.4.1 Kriterien

Priorität	Kriterium	Media Wiki	Kommentar
1	Open Source	x	
1	Suchfunktion	x	Ausweitung auf semantische Suche möglich
1	Einfache Navigation	x	
1	Ablage von Dateien	x	
2	70-150 User	größer	Die Nutzeranzahl ist unbegrenzt
2	Versionierung	x	
2	Speicher Datenmengen <20.000	x	
3	Verbreitung/Reife	x	Media Wiki wurde für die Plattform Wikipedia entwickelt
3	einfache Implementierung	/	Die Servergröße ist wichtig, kleinere Server haben mehr Probleme
3	Datengröße	x	
4	Handling von Anhängen	/	
4	Lizenzmodell	x	
4	Berechtigungssystem	x	
5	Gestaltungsmöglichkeiten	/	Es ist keine individuelle Gestaltung möglich
5	Einsetzbarkeit	x	
5	Autorensystem	x	

Tabelle 7: Kriterienkatalog von Media Wiki

Die Plattform ist unter der GNU General Public License (GPL) lizenziert. Das heißt der Source Code von Media Wiki ist für alle frei einsehbar. Somit ist das Kriterium der Priorität eins, welches Open Source voraussetzt hiermit gegeben. [185]

Media Wiki besitzt außerdem eine Suchfunktion, welche im Kriterienkatalog (siehe Tabelle 7) auch unter Priorität eins fällt. Sie ist leicht auf der rechten Seite von Media Wiki mit einem Lupensymbol und einem Suchfeld zu finden. Wenn eine Seite nicht gefunden wird, wird automatisch eine Reihe von Artikeln mit ähnlichem Inhalt vorgeschlagen. Jedoch gibt es hier auch einige Schwachpunkte. Bei Media Wiki wird nach jedem Wort einzeln gesucht,

[185] Vgl. Media Wiki (2014e)

unabhängig ob in Anführungszeichen oder mit Großbuchstaben.[186] Hier gibt es jedoch eine freie Erweiterung, welche Semantic Media Wiki (SMW) heißt. Bei SMW erscheint die Suchfunktion in Form einer Abfragesprache, welche es den Benutzern ermöglicht, auf die Informationen des SMW zuzugreifen.[187] Die Informationen werden also nicht mehr im Volltext gesucht, sondern können direkt von den Nutzern abgefragt werden. [188] Die Inhalte können weiterhin durchsucht, gebrowst, ausgewertet und mit anderen Nutzern geteilt werden. Die Erweiterung besteht darin, dass SMW semantische Annotationen („Hilfsmittel um Artikel nach bestimmten Kriterien zu klassifizieren“[189]) hinzufügt und so die Möglichkeiten des Semantischen Webs in das Wiki überbringt. Ein semantisches Web bringt verwandte Lösungen in Beziehung, um mögliche Ansätze zu bestimmen. Das bedeutet Inhalte sollen nicht nur eine Bedeutung haben, sondern auch in Beziehung zu anderen Bedeutungen stehen. Es sollen somit hierarchische Klassen oder Ausschlusskriterien gebildet werden.[190] Ein einfaches Beispiel ist: „Ein LKW ist ein Auto, aber weder PKW noch Geländewagen.“[191] Es lässt sich hier deutlich erkennen, dass bei Semantic Media Wiki die Suchfunktion sehr stark ausgeprägt und sehr genau ist.

Ein weiterer wichtiger Punkt im vorliegenden Kriterienkatalog ist die Navigation durch die einzelnen Funktionen durch das Tool. Bei Media Wiki ist diese strukturiert und für die Nutzer verständlich. Zu Beginn sehen die anonymen Nutzer ein Anmeldefeld. Bereits eingeloggte Nutzer haben eine Auswahl an persönlichen Links, welche einen Link zu der eigenen Benutzerseite beinhaltet. Eine Seitenleiste ermöglicht den Zugang zu den wichtigen Wiki Seiten. Diese sind beispielsweise die letzten Änderungen oder auch „Datei hochladen“. Außerdem beinhalten alle Seiten Links, welche auf Spezialseiten verleiten. Spezialseiten sind „solche Seiten, die aktuelle Informationen über das Wiki enthalten und beim Aufrufen erstellt werden oder für spezielle administrative Aktionen.“[192]

Des Weiteren gibt es bei Media Wiki sogenannte Artikelreiter. Siehe Abbildung 13. Diese Reiter sehen unterschiedlich aus, abhängig von dem Status der Benutzer (angemeldet, alle Benutzer, Administratoren). Es wird deutlich, dass die Nutzung von Media Wiki auch ohne große Schulungen möglich ist.

[186] Vgl. Media Wiki (2014b)
[187] Vgl. Semantic Media Wiki (2012b)
[188] Vgl. Hansch D./Schnurr H./Pissierssens P.(2009), S.211
[189] Semantic Media Wiki (2012a)
[190] Vgl. Merschmann H.(2008)
[191] Merschmann H. (2008)
[192] Media Wiki (2014g)



Abb. 13: Artikelreiter zur Navigation bei Media Wiki Administratorensicht[193]

Wichtig für eine solche Plattform ist auch das Hochladen von unterschiedlichen Quellen. Diese Funktion ist im Kriterienkatalog ebenfalls mit der Priorität eins behaftet. Bei Media Wiki hängt das Hochladen von drei Faktoren ab.[194]

1. Der Parameter $wgEnableUploads muss auf true im Datei LocalSettings.php gesetzt werden
2. Der Dateityp muss zulässig sein
3. Der Anwender muss angemeldet sein

Die Administratoren können verschiedene Dateien erlauben, welche hochgeladen werden dürfen. Dies muss in den LocalSettings.php festgehalten werden. Wichtig ist es einen Virusscan für hochgeladene Dateien zu ermöglichen. Diese vermindern das Risiko von ungewolltem Datenverlust enorm.

Es gibt jedoch auch eine sogenannte Black List. Diese enthält alle Dateien, welche in keiner Weise auf Media Wiki hochgeladen werden können.[195] Siehe Abbildung 14. Somit ist die Priorität eins abgeschlossen und Media Wiki erfüllt alle diese.

[193] Mit Änderungen entnommen aus: Media Wiki (2014g)
[194] Vgl. Media Wiki (2014d)
[195] Vgl. Media Wiki (2014a)

Files with these extensions will never be allowed as uploads if `$wgCheckFileExtensions` is set to true.

This is the default value:

```
$wgFileBlacklist = array(
  # HTML may contain cookie-stealing JavaScript and web bugs
  'html', 'htm', 'js', 'jsb', 'mhtml', 'mht', 'xhtml', 'xht',
  # PHP scripts may execute arbitrary code on the server
  'php', 'phtml', 'php3', 'php4', 'php5', 'phps',
  # Other types that may be interpreted by some servers
  'shtml', 'jhtml', 'pl', 'py', 'cgi',
  # May contain harmful executables for Windows victims
  'exe', 'scr', 'dll', 'msi', 'vbs', 'bat', 'com', 'pif', 'cmd', 'vxd', 'cpl' );
```

Abb. 14: Black List[196]

Anschließend wird sich mit Priorität zwei beschäftigt. Hierunter fällt unter anderem die Begrenzung der Benutzer des Tools. Diese Begrenzung gibt es bei Media Wiki nicht, somit wird das Kriterium „70-150 User" nicht eingeschränkt und es können deutlich mehr Nutzer das Tool verwenden, denn es wurde für offene Inhalte und Server-Farmen mit Millionen von Seitenzugriffen pro Tag entwickelt.[197] Denkt man nur ein weiteres Mal an Wikipedia, so wird einem das Ausmaß, für welches Media Wiki entwickelt wurde, klarer.

Des Weiteren ist die Versionierung ein wichtiger Punkt bei Priorität zwei. Bei Media Wiki kann eine alte Datei durch eine neue Version ersetzt werden. Somit entsteht eine sogenannte Bildhistorie. Diese wird im zugehörigen Artikel angegeben. Sobald jedoch eine Datei gelöscht wird, ist diese nicht wiederherstellbar. Also ist auch dieses Kriterium erfüllt.[198]

Die Höhe der speicherbaren Datenmenge hat ebenfalls Priorität zwei erhalten. Bei Media Wiki ist es per Standard erlaubt von einer Webanwendung nur 30 MB hoch zuladen. Dies kann man jedoch umgehen, indem man die maximale Dateigröße ersetzt. Das heißt die Administratoren können eigene Grenzen nach oben in Bezug auf die Höhe der hochgeladenen Dateien setzen. Dies führt zum Ende der Priorität zwei und auch hier schlägt sich Media Wiki gut.[199]

Priorität drei des Kriterienkatalogs beschäftigt sich neben anderen Kriterien auch mit der Verbreitung und der Reife. Media Wiki ist stark verbreitet, da das Tool ursprünglich für Wi-

[196] Enthalten in: Media Wiki (2014c)
[197] Vgl. Media Wiki (2014e)
[198] Vgl. Media Wiki (2014e)
[199] Vgl. Media Wiki (2014a)

kipedia entwickelt wurde.[200] Somit hat der Großteil der Menschen schon einmal mit Media Wiki gearbeitet.

Auch beschäftigt sich Priorität drei mit der einfachen Implementierung der Software. Die Voraussetzungen für die Nutzung von Media Wiki sind ein Webserver und eine Datenbank. Sobald diese vorhanden sind, kann die Software installiert werden.[201] Media Wiki kann auf Windows 7/8, Windows Vista und XP, Mac OS und Linux installiert werden. Jedoch gibt es hier einige Schwachstellen, welche nicht außer Acht gelassen werden dürfen. Vor allem für kleinere Sever ist Media Wiki nicht optimal, da es für Plattformen wie Wikipedia entwickelt wurde. Bei kleineren Servern sind oft Plattenplatz oder RAM größere Einschränkungen als die Bandbreite.[202]

Das letzte Kriterium bei Priorität drei ist die Datengröße, welche hochgeladen werden kann. Standardmäßig erlaubt PHP das Hochladen von Dateien mit maximal 2 MB Größe. Dies kann jedoch umgangen werden, wenn man zwei Parameter in der PHP Konfiguration verändert.[203]

Der Umgang mit Anhängen ist ein weiteres Kriterium. Dieses ist mit Priorität vier behaftet. Bei Media Wiki werden immer verschiedene Einträge erzeugt, sobald eine Datei hochgeladen wird.

1. Es wird ein Artikel im Namensraum Bild mit dem exaktem Namen der Datei erstellt, gespeichert und verwaltet
2. Die Datei wird in einem Ordner auf dem (Unix-)Server gespeichert
3. Sobald die Datei eine bestimmte Größe überschreitet, wird zusätzlich eine kleinere Version erzeugt und in einen separaten Ordner angelegt. Jede dieser Dateien hat einen eigenen Ordner mit verkleinerten Versionen

Das bedeutet, dass alle Dateien in einem eigenen Ordner gespeichert werden. Des Weiteren werden sogenannte Thumbnails erstellt. Das sind kleine Bildversionen für die hochgeladenen Dateien. Wichtig ist jedoch, dass es, sofern es sich um keine Bilddatei handelt, mit einem Datei Icon angezeigt wird.[204]

[200] Vgl. Media Wiki (2014e)
[201] Vgl. Carl D./ Eidenberger H./ Ludewig M./Mintert S./Schulz C./Spanneberg B./Völkl G.(2008)
[202] Vgl. Media Wiki (2014e)
[203] Vgl. Media Wiki (2014a)
[204] Vgl. Media Wiki (2014d)

Im Kriterienkatalog mit der Priorität vier behaftet ist auch das Lizenzmodell. Auch dieses Kriterium erfüllt Media Wiki, denn zusätzlich zu dem Standardmodell können einige Erweiterungen zu Media Wiki hinzuinstalliert werden, wie beispielsweise Semantic Media Wiki (bereits oben erwähnt). Auch SMW ist kostenfrei.

Ein weiteres Kriterium mit Priorität vier ist das Berechtigungssystem. In Media Wiki gibt es einige verschiedene Berechtigungen:

Alle Benutzer:

Jeder darf Seiten anlegen und bearbeiten – der Benutzer muss nicht angemeldet sein. Dies lässt sich darauf zurückführen, dass Media Wiki für offene Inhalte entwickelt wurde.[205] (siehe Wikipedia). Die Beiträge nicht angemeldeter Nutzer erscheinen in der Versionsgeschichte unter der IP-Adresse, welche beim Einwählen ins Internet zugewiesen wird. Es ist somit nicht möglich den Verfasser eindeutig zu identifizieren.

Angemeldeter Benutzer:

Zusätzlich zu den Rechten aller Benutzer können die User Seiten verschieben und Dateien hochladen. In der Versionsgeschichte des Media Wiki erscheint ihr Benutzername. Des Weiteren erhält jeder User eine eigene Seite im Benutzernamensraum. Dort kann der angemeldete Benutzer Seiten auf seine Beobachtungsliste setzen. Es besteht die Möglichkeit eine Liste aller angemeldeten Nutzer zu erhalten.

Administrator (sysop):

Den Administratoren obliegt das Recht verschiedene Seiten zu schützen, diese zu bearbeiten, löschen und die Wiederherstellung der gelöschten Seiten. Außerdem haben Administratoren die Macht andere Benutzer bzw. IPs zu sperren und diese Sperren auch wieder aufzuheben. Wichtig ist, dass die gesperrten User oder IPs trotzdem die Seiten lesen können, jedoch nichts mehr verändern.

Bürokrat (bureaucrat):

Bürokraten können allen anderen Benutzergruppen die Rechte erteilen und entziehen.

Bot:

Ein sogenannter Bot muss mit Hilfe eines Programmes oder Skriptes häufig auftretende Aufgaben erledigen, wie beispielsweise die Tippfehler hochgeladener Dokumente korrigieren.[206]

[205] Vgl. Media Wiki (2014e)
[206] Vgl. Wikibooks (2013)

Durch die verschiedenen Berechtigungen liegt eine gewisse Struktur bei Media Wiki vor und man ermöglicht den Administratoren die Kontrolle über das eigene Media Wiki. Wichtig ist jedoch auch, dass durch die Tatsache, dass jeder die Einträge lesen kann das eigene Media Wiki nicht privat ist.

Die letzte Priorität im Kriterienkatalog ist Priorität fünf. Hierunter steht unter anderem die Möglichkeit der unterschiedlichen Gestaltung von Media Wiki. Hier bietet die Software keine große Personalisierung der Webseiten für Layout und Design.[207]

Ein weiteres Kriterium des vorliegenden Kriterienkatalogs ist die Einsetzbarkeit. Nachdem Media Wiki über einen Webserver läuft, ist es kein Problem das Tool über ein mobiles Endgerät zu öffnen und zu verwenden.

Auch muss in diesem Zusammenhang das das Autorensystem beachtet werden. Um die volle Übersichtlichkeit eines Tools zu garantieren ist es notwendig zu sehen, welcher Benutzer welche Datei in Media Wiki hochgeladen hat. Um eine Übersicht über die verschiedenen hochgeladenen Dateien des Tools zu bekommen, kann sich der Nutzer auf verschiedene Spezialseiten begeben. Hier werden jedoch nur die Nutzer angegeben, welche auch angemeldet sind:

1. *Gallery of new files*
   ➔ Hier kann der User die neuen hochgeladenen Dateien einsehen
2. *File list*
   ➔ Diese Spezialseite zeigt alle Dateien mit den zugehörigen Usernamen auf
3. *Unused files*
   ➔ Diese Seite hilft nutzlose und bereits vergessene Dateien zu finden[208]

Hier wird schnell deutlich, dass es für die Benutzer möglich ist, verschiedene Informationen bezüglich der hochgeladenen Dateien zu finden.

5.4.2 Zusatzfunktionen

Eine hilfreiche Zusatzfunktion von Media Wiki ist Semantic Media Wiki, wie bereits erwähnt. Hier kann man die Suche mit semantischen Annotationen erweitern. Des Weiteren können die Artikel über *Interlanguage-Links* in mehreren verschiedenen Sprachen verknüpft werden. Dies ist vor allem bei internationaler Anwendung von großem Vorteil. Außerdem gibt es so-

[207] Vgl. Wikipedia (2015)
[208] Vgl. Media Wiki (2014f)

genannte *Interwiki-Links.* Diese verweisen zu anderen Wiki Projekten. Beispielsweise zu Artikeln anderer Sprachversionen oder zu unterschiedlichen Projekten.[209]

Zusammenfassend lässt sich sagen, dass Media Wiki ein sehr erfolgreiches Tool ist. Vor allem die Ergänzung mit der semantischen Suche, erleichtert es Unternehmen stark wichtige Dateien wieder zu finden. Nicht außer Acht gelassen werden darf jedoch die Tatsache, dass es für große Serverfarmen entwickelt wurde und auf kleineren Servern nicht optimal implementiert werden kann. Eine weitere Schwachstelle bei Media Wiki ist die Tatsache, dass es keine Privatsphäre gibt. Jeder kann die Inhalte des Media Wikis lesen.

5.5 Wordpress

WordPress ist ein Blog und das weltweit meist genutzte Content Management System (CMS).[210] Die Software ist ein Open-Source Projekt einer sehr großen Community, d.h. es gibt kein Unternehmen dahinter. Mehrere Unternehmen haben sich gegründet, um die Software als Teil der Community noch besser zu machen, die auch verschiedene Plug-Ins anbieten und kostenpflichtige Services oder Support anbieten. Im Folgenden wird nun das Basis-Produkt auf die zuvor in Kapitel 3.4 definierten Kriterien untersucht.

5.5.1 Kriterien

Priorität	Kriterium	WordPress	Kommentar
1	Open Source	x	
1	Suchfunktion	x	Zusätzliche Plugins vorhanden
1	Einfache Navigation	x	Editor
1	Ablage von Dateien	x	Bilder, Textdokumente, Videos, PDFs, Microsoft Office Dokumente
2	70-150 User	?	abhängig vom Server
2	Versionierung	x	Nicht von Dateien, nur Artikel
2	Speicher Datenmengen <20.000	?	anzunehmen
3	Verbreitung/Reife	x	weltweit, meist genutztes CMS
3	einfache Implementierung	/	langwierige Definition der Struktur etc
3	Datengröße	x	Bis zu 8MB, erweiterbar durch PlugIn

[209] Vgl. Wikipedia (2014b)
[210] Vgl. Inpsyde GmbH (o.J.a)

4	Handling von Anhängen	?	Dateien werden auf dem Server gespeichert
4	Lizenzmodell	x	GPL 2
4	Berechtigungssystem	x	5 Benutzerrollen
5	Gestaltungsmöglichkeiten	x	Designanpassungen durch Themes, Design-Forum
5	Einsetzbarkeit	x	Apps, Login via Webbrowser
5	Autorensystem	x	

Tabelle 8: Kriterienkatalog von WordPress

WordPress ist, wie zuvor erwähnt, eine Open-Source Lösung.[211] Da die Community dieses Projektes sehr groß ist, ist die Qualität der Software sehr gut und es gibt kaum Fehler, die nicht bald behoben werden. Sobald das System auf einem eigenen Server installiert wird, gibt es auch keine Kosten für das Hosting.

Eine Suchfunktion ist in der Basis-Version vorhanden und kann durch verschiedene Plug-Ins noch verbessert und angepasst werden.[212]

Durch den Editor ist die Navigation durch das System sehr einfach. Man kann Texte einfach eingeben und dazugehörige Dokumente hochladen. Eine Struktur kann durch ein Theme-Template oder ein eigen erstelltes Theme vorgegeben und definiert werden, wodurch die Navigation noch einfacher wird.[213]

Das letzte Kriterium der ersten Priorität ist das Ablegen von verschiedenen Dateitypen. Es ist möglich die Dateiformate .pdf, .jpg, .jpeg, .png, .gif, .doc, .ppt, .pptx, xls, .zip .mp3, .wav, .wmv etc hochzuladen.[214] Somit sind alle Kriterien der ersten Priorität erfüllt. Im Folgenden werden die Kriterien der zweiten Priorität untersucht.

WordPress hat fünf verschiedene Benutzerrollen. Es gibt Admins, die alle Rechte haben. Diese bestimmen Themes etc. und ist für die Benutzerverwaltung verantwortlich. Redakteure können ebenfalls Artikel bearbeiten und veröffentlichen, Dateien hochladen und kommentieren. Autoren können nur die eigenen Beiträge bearbeiten und hochladen. Mitarbeiter können nur eigene Beiträge bearbeiten. Wenn diese einen neuen Artikel veröffentlichen wollen, brauchen sie die Erlaubnis des Admins. Follower können nur lesen und nichts bearbeiten

[211] Vgl. Inpsyde GmbH (o.J.b)
[212] Vgl. Schwerthalter, R. (2014)
[213] Vgl. Inpsyde GmbH (o.J.a)
[214] Vgl. Automattic (o.J.a)

oder veröffentlichen. Die .Versicherung sollte einen bis zwei Admins bestimmen, die auch die Struktur festlegen. Alle anderen Mitarbeiter der Abteilung Anwendungsentwicklung sollten Redakteure sein, um auf alle notwendigen Funktionen zugreifen zu können.[215] Wie viele Redakteure ein WordPress Blog haben kann, konnte im Rahmen der Recherche dieser Projektarbeit nicht erarbeitet werden. Die Anzahl der möglichen User ist abhängig vom Server. Aufgrund der vielen Blogs, die auf Basis von WordPress laufen ist aber anzunehmen, dass mindestens 150 Redakteure möglich sind.

Das Kriterium der Versionierung ist bei allen Artikeln vorhanden. Wie in Abbildung 15 zu sehen ist, wird angezeigt wer wann was geändert hat. Die Wiederherstellung von älteren Artikeln ist ebenso möglich.[216] Jedoch gibt es keine Versionierung von hochgeladenen Dateien.

Artikelüberarbeitungen

17.09.2008 17:16 [Automatische Speicherung] von Thordis
17.09.2008 17:11 von Thordis
11.09.2008 10:05 von Thordis
08.09.2008 22:28 von Thordis
08.09.2008 20:57 von Thordis
08.09.2008 20:57 von Thordis
08.09.2008 20:52 von Thordis
08.09.2008 20:49 von Thordis

Abb. 15: Versionierung WordPress[217]

Die maximale Anzahl der Daten, die bei WordPress gespeichert werden können bzw. der Dateien, die hochgeladen werden können, ist im Rahmen der Recherche nicht herausgefunden worden.

Somit sind alle Kriterien der Priorität zwei erfüllt. Im Folgenden werden alle Kriterien der Priorität drei untersucht.

Die Verbreitung und Reife der Open-Source Software WordPress ist kaum zu übertreffen. 2003 starteten Matt Mullenweg und Mike Little das Projekt, welches mittlerweile das größte CMS System der Welt darstellt. Mit über 66 Millionen Installationen hat WordPress einen Marktanteil unter den CMS Systemen von 66%. 15% aller Websites weltweit basieren auf

[215] Vgl. Automattic (o.J.b)
[216] Vgl. perun.net webwork gmbh (2014)
[217] Mit Änderungen enthalten in: perun.net webwork gmbh (2014)

WordPress, unter anderem die Unternehmens-Seiten von Intel, Yahoo, ZDF, EBay, adidas, VW, SAP.[218]

Die Implementierung des Systems ist schwerer als andere Systeme, da es durch die individuelle Nutzbarkeit der Software (als Blog, CMS oder Website) viele Eigenschaften zu definieren gibt und zunächst eine Struktur aufgebaut werden muss. Benötigt wird ein Webspace, der folgende Voraussetzungen erfüllt:

1. 1. PHP-Version 5.2.4 oder höher
2. 2. MySQL-Version 5.0 oder höher
3. 3. Das Apache mod_rewrite Modul, für „schönere URLs“

Es wird ein Linux-Server mit Apache empfohlen.[219] Danach kann die einfache Installation sofort beginnen. Die Definition der Struktur, der Themes und die Installation eventueller PlugIns nimmt dann Zeit in Anspruch.

Die maximale Dateiuploadgröße beträgt bei WordPress 8 MB. Diese kann man jedoch durch verschiedene PlugIns bei Bedarf erhöhen.[220]

Die Kriterien der Priorität drei sind alle weitgehend erfüllt, wobei WordPress bei der einfachen Implementierung nicht überzeugen kann. Nun werden die Kriterien der Priorität vier untersucht.

Anhänge können ganz einfach zu einem dazugehörigen Artikel hochgeladen werden. Bilder können sogar über Drag&Drop in den Artikel hinzugefügt werden.[221] Eine weitere Möglichkeit wäre die Dokumente als Artikel abzuspeichern, damit die Texte bzw. Inhalte des Dokuments sofort sichtbar sind, ohne es per Mausklick noch öffnen zu müssen. Dann würde auch die Versionierung immer greifen. Wie die Speicherung der Daten exakt erfolgt, konnte im Rahmen der Projektarbeit nicht herausgefunden werden. Es ist aber anzunehmen, dass WordPress Anhänge in der dahinter liegenden Datenbank auf dem Server speichert, auf dem das System läuft. Die Aufrufe erfolgen.

Das Lizenzmodell ist GPL2.[222]

[218] Vgl. Inpsyde GmbH (2013)
[219] Vgl. Inpsyde GmbH (o.J.c)
[220] Vgl. Inpsyde GmbH (o.J.d)
[221] Vgl. Inpsyde GmbH (o.J.a)
[222] Vgl. Inpsyde GmbH (o.J.b)

Wie bereits in einem vorherigen Kriterium „Anzahl der User“ oben erläutert, gibt es fünf Benutzerrollen. Die Administratoren müssen diese Rollen vergeben und verwalten.[223] Außerdem ist es möglich, die Artikel mit einem Passwort zu schützen, falls man als Redakteur nur ganz bestimmten Personen Zugriff geben möchte.[224]

Die Kriterien der Priorität vier sind alle erfüllt, es folgt nun die Untersuchung der letzten Priorität fünf.

Die Gestaltungsmöglichkeiten bei der Software WordPress sind sehr vielfältig. Durch Themes kann man die Struktur der Website komplett selbst bestimmen und durch PlugIns sogar noch durch weitere Funktionalitäten erweitern. Auch Design-Forums sorgen für die Individualisierbarkeit der Software.[225]

Die Einsetzbarkeit des Systems ist sehr gut. Der Zugriff zum Verfassen von Artikeln erfolgt über das Einloggen mit User-Name und Passwort über jeden Webbrowser oder eine App. Die Apps sind Tablet und Smartphone fähig und für beinahe alle Betriebssysteme programmiert.[226]

Die Funktionalität des Autorensystems ist vorhanden. In der Historie wird unter jedem Artikel angezeigt wer wann etwas verändert hat. Beim Anwählen des User-Namens sind alle Beiträge und Artikel des Users auf einen Blick zu sehen.

5.5.2 Zusatzfunktionen

Zu den Artikeln können auch Kommentare und Links hinzugefügt werden.[227] Da das Angebot von Erweiterungen und Zusatzfunktionen von WordPress durch die vielen verschiedenen kostenlosen PlugIns sehr vielfältig und groß ist, werden in dieser Projektarbeit nicht noch mehr PlugIns beschrieben. Ein offizielles Plugin Verzeichnis ist auf der WordPress Deutschland Webpage zu finden.[228]

WordPress hat alle Kriterien erfüllt. Die Implementierung und Strukturierung zu Beginn ist jedoch langwierig und aufwendig, damit der Blog als Dokumentenverwaltungssystem genutzt werden kann.

[223] Vgl. Automattic (o.J.b)
[224] Vgl. Inpsyde GmbH (o.J.a)
[225] Vgl. Inpsyde GmbH (o.J.e)
[226] Vgl. Inpsyde GmbH (o.J.f)
[227] Vgl. Inpsyde GmbH (o.J.a)
[228] Vgl. Inpsyde GmbH (o.J.g)

6 Schluss

6.1 Zusammenfassung der Ergebnisse

Nachdem die Top fünf Wikis analysiert wurden, werden sie in der nachstehenden Tabelle anhand der Kriterien aus dem Kriterienkatalog miteinander verglichen.

Die untersuchten Dokumentationsplattformen unterscheiden sich in ihrer Art: MediaWiki und DokuWiki sind Wiki-Software, Wordpress ist eine Software, die hauptsächlich zur Erstellung von Webseiten und Blogs gedacht ist, Agorum und Alfresco sind Systeme zum Managen von Dokumenten.

Das erste Kriterium einer Open Source Software wird von allen Dokumentationsplattformen teilweise erfüllt, denn jede der Dokumentationsplattformen ist quelloffen, das heißt der Quellcode ist verfügbar. Allerdings ist Alfresco das einzige Tool welches kostenpflichtig ist, die anderen sind weitgehend kostenlos. Regelmäßige Updates und Zusatzfunktionen für Agorum kosten ebenfalls einen geringen Beitrag.
DokuWiki und MediaWiki sind in manchen Bereichen sehr ähnlich aufgebaut. Allerdings ist DokuWiki in einigen Bereichen ein wenig anwenderfreundlicher, beispielsweise bei der Installation. MediaWiki zeigt sich für geschlossene Organisationen als ungeeignet, da Jeder in das Wiki einsehen kann. Ein Vorteil von MediaWiki gegenüber DokuWiki stellt die verwendete Datenbank dar, auf die DokuWiki komplett verzichtet. Dadurch ist die Skalierbarkeit höher, was sich bespielsweise auf die Geschwindigkeit der Suche auswirkt.
Wordpress ist das größte CMS System der Welt. Durch die extrem große Community werden Fehler schnell behoben, was die Software qualitativ sehr hochwertig macht. Ein Nachteil von Wordpress gegenüber den anderen Programmen ist, dass die Software besonders für den Einsatz als Blog geeignet ist, was einen sehr hohen Aufwand mit sich zieht, um aus Wordpress ein Dokumentenverwaltungssystem zu machen.
Der Vorteil von Alfresco gegenüber den anderen Systemen ist der angebotene weltweite Kundensupport. Außerdem ist das System ähnlich aufgebaut wie ein virtuelles externes Netzwerk, auf das man Dokumente per Drag and Drop hochladen kann, und dadurch besonders anwenderfreundlich.
Ähnlich wie bei Alfresco ist Agorum ebenfalls in einer Ordnerstruktur aufgebaut, sodass Nutzer eine bekannte Umgebung haben. Einige Funktionen hat Agorum vergleichbar aufgebaut wie bei einem Wiki. So kann man zum Beispiel Artikel kommentieren oder verändern. Es ist auch möglich ein Wiki an Agorum anzubinden.

Zusammenfassend lässt sich sagen, dass die untersuchten Plattformen sich zwar in einigen Punkten ähneln, dennoch einige Unterschiede zu erkennen sind. Alle fünf Tools sind sehr rennomiert und werden häufig eingesetzt, da sie eine Vielzahl von Funktionen bieten.

Kriterium	Agorum Core Open	Alfresco	DokuWiki	MediaWiki	Wordpress
Open Source	√	√	√	√	√
Suchfunktion	sehr gut, aber kein Highlighting in Open Source Version	Volltextsuche	Index-basierte Volltextsuche	Ausweitung auf semantische Suche möglich	vorhanden und Plugins
Einfache Navigation	OneClick zum Öffnen		Breadcrumbs-Navigation		einfach durch den Editor
Ablage von Dateien	Ordner, Dokumente, E-Mails, Wikis, Foren, Termine, Benutzer, Gruppen, Eigene Objekttypen		Anhänge (Bilder, Videos, pdfs, …)		Bilder, Textdokumente, Videos, PDFs, Microsoft Office Dokumente
70-150 User	Keine Begrenzung	Keine Begrenzung	Keine Begrenzung	Keine Begrenzung	nicht sicher, abhängig vom Server
Versionierung	Automatische Historie von Dokumenten-Änderungen		Ältere Versionen werden aufbewahrt		von Artikeln aber keine Dateien
Speicher Datenmengen <20.000	Keine Begrenzung	größere Datenmengen sind möglich	keine Begrenzung		
Verbreitung/Reife	deutsch, 10000 Downloads, Produkt seit 2008, Unternehmen seit 1998, gute Referenzen, Schwäbisch Hall		Sehr verbreitet, regelmäßige Releases	Media Wiki wurde für die Plattform Wikipedia entwickelt	weltweit, meist genutztes CMS
einfache Implementierung	einfache Installation auf einem Server Implementierung in den Windows Explorer, Zugriff über Internet Browser	Webanwendung, einmalige Installation genügt	Anleitungen verfügbar	Die Servergröße ist wichtig, kleinere Server haben mehr Probleme	Einfache Installation, langwierige Definition der Struktur etc
Datengröße	Keine Begrenzung	Keine Begrenzung	Keine Begrenzung		Bis zu 8MB, erweiterbar durch PlugIn
Handling von Anhängen	Dokumentenverwaltungssystem, Speicherort nicht exakt bekannt, Server ist anzunehmen		Durchsuchen von Anhängen		Dateien werden auf dem Server gespeichert
Lizenzmodell	GPL v2	genaue Preisauskunft vom Hersteller nicht möglich	GPL v2	GPL v2	GPL v2
Berechtigungssystem	Gruppen, Benutzer, Access Control Lists,		Zugriffskontrolle durch Listen		es gibt 5 Benutzerrollen
Gestaltungsmöglichkeiten	Ordner-basiert, eigene Objekttypen mit eigenen Attributen, Meta-Daten, Integration eigener Masken		Vielfältig – Templates, Plugins	Es ist keine individuelle Gestaltung möglich	Designanpassungen durch Themes, Design-Forum
Einsetzbarkeit	Laufwerk oder über einen Webbrowser (Smartphone oder Tablet mit HTML5-fähigen Browser)		Webanwendung		Apps, Login via Webbrowser
Autorensystem	Historie wird angezeigt		Autoren von Änderungen sichtbar/ Anhänge?		Historie wird angezeigt

Tabelle 9: Zusammenfassung der Ergebnisse

6.2 Finale Empfehlung mit Hilfe einer Nutzwertanalyse

Anhand der Übersicht (siehe Kapitel 6.1), in der alle fünf Tools miteinander verglichen werden, kann die Liste der in Frage kommenden Tools für die .Versicherung von fünf auf drei reduziert werden. Die drei favorisierten Tools sind Agorum, Alfresco und DokuWiki.

Die Tools MediaWiki und WordPress wurden aussortiert. WordPress wurde nicht in die engere Wahl gezogen, da es in der eigentlichen Funktion ein Blog ist und es zeitaufwändig wäre, diesen zu einem Dokumentationsverwaltungstool anzupassen. MediaWiki wurde abgelehnt, da Jeder in das Wiki einsehen kann. Da das Dokumentationstool vor allem zur Ablage von personenbezogenen Informationen gedacht ist und die Datensicherheit eine immer größere Rolle spielt, ist dieses Tool ungeeignet für den Verwendungszweck der .Versicherung .

Die drei verbliebenen Tools DokuWiki, Alfresco und MediaWiki werden einer Nutzwertanalyse unterzogen, um eine finale Entscheidung treffen zu können, welches Tool für die Zwecke der .Versicherung am besten geeignet ist. In der unten abgebildeten Nutzwertanalyse zeigt die zweite Spalte die Gewichtung der einzelnen Kriterien. Entsprechend der Prioritätenwahl bei der Erstellung des Kriterienkatalogs wurde bei der Entwicklung der Nutzwertanalyse die Gewichtung verteilt: die Kriterien mit der höchsten Priorität erhielten die Gewichtung fünf, welche für „unverzichtbar“ steht. Die restlichen Prioritätsgruppen wurden in der Reihenfolge mit absteigenden Werten gewichtet. In dem nächsten Schritt wurden alle drei Favoritentools unabhängig von der Gewichtung anhand einer Skala von eins („nicht erfüllt“) bis zehn („komplett erfüllt“) auf ihren Erfüllungsgrad bewertet. Hierbei ist es wünschenswert, dass die Tools bei den wichtigsten Kriterien auch den höchsten Erfüllungsgrad aufweisen. In dem dritten Schritt wurden alle Erfüllungswerte mit der jeweiligen zuvor festgelegten Gewichtung multipliziert. Die jeweiligen Ergebnisse pro Tool pro Spalte können in den Spalten „Ergebnisse“ nachvollzogen werden. In dem vierten und letzten Schritt wurden die Ergebnisse jeder Ergebnisspalte aufsummiert.

Es ist auffällig, dass alle drei Tools in dem Erfüllungsgrad der drei wichtigsten Kriteriengruppen (Kriterien eins bis zehn) sehr ähnlich abschneiden und deswegen auch keine klare Führung ersichtlich wird.
Das Tool DokuWiki ist mit 477 Gesamtpunkten der knappe Sieger des Rankings. Dicht hinter DokuWiki, mit einer Gesamtzahl von 472 Punkten, liegt Alfresco. Agorum liegt mit 460 Punkten auf Platz drei. Dementsprechend ist es schwierig, anhand der nah aneinander liegenden Ergebnisse der Nutzwertanalyse, eine finale Entscheidung zu treffen.

Grundsätzlich kann jedoch bei der Entscheidung berücksichtigt werden, dass Unternehmen, die sich für den Einsatz von Open Source Software entscheiden, hauptsächlich aus Kostengründen handeln. Bei dem Kauf kommerzieller Software sind der Kaufpreis und die Einführungskosten nicht die einzigen Kosten. TCO-Analysen (Total Cost of Ownership) listen sämtliche Kosten auf, die im Laufe des Lebenszyklus von eingeführter Software von der Einführung bis hin zu der Außerbetriebnahme anfallen[229] und zeigen, dass der Betrieb von Software ein erheblicher Faktor ist, der bei der Anschaffung von IT einbezogen werden sollte. Somit handeln Unternehmen, die die Einführung von Open Source Software bevorzugen, aus Kostengründen heraus: sie möchten Kosten für anfallende Updates und/oder Support vermeiden.

Wird bei der finalen Entscheidungsfindung berücksichtigt, dass die Einführung von Alfresco mit kostenpflichtigem Support verbunden ist, kann dieses Produkt ausgeschlossen werden.
Bei dem Einsatz von Agorum ist festzuhalten, dass auf freiwilliger Basis kostenpflichtige Updates anfallen. Hierbei muss vor allem der Faktor Aktualität in die Entscheidung miteinbezogen werden: Gibt es Performanceeinbußungen, wenn die angebotenen Updates nicht in Anspruch genommen werden? Ein weiterer Faktor, der im direkten Vergleich zwischen Agorum und DokuWiki für DokuWiki spricht, ist der Verbreitungsgrad. Agorum wurde bisher nur deutschlandweit eingesetzt, DokuWiki dagegegen im internationalen Bereich, was für einen höheren Reifegrad spricht.

Aufgrund des Ausschlusses von Agorum und Alfresco wird der Anwendungsentwicklung der .Versicherung das Tool DokuWiki als neues Dokumentenablagesystem empfohlen, denn es erfüllt alle Anforderungen an das Tool und ist darüber hinaus mit keinerlei Kosten verbunden, weder bei der Einführung des Tools noch bei dem Betrieb.

[229] Vgl. Brem, geb. Krämer, S. (2009), S.6f.

Kriterium	Gewichtung	Erfüllungsgrad Agorum	Ergebnis	Erfüllungsgrad Doku Wiki	Ergebnis	Erfüllungsgrad Alfresco	Ergebnis
Open Source	**5**	10	50	10	50	10	50
Suchfunktion	**5**	8	40	8	40	8	40
Einfache Navigation	**5**	10	50	9	45	10	50
Ablage von Dateien	**5**	10	50	10	50	10	50
70-150 User	**4**	10	40	10	40	10	40
Versionierung	**4**	10	40	10	40	10	40
Speicherbare Datenmengen <20.000	**4**	10	40	10	40	10	40
Verbreitung/Reife	**3**	8	24	10	30	10	30
einfache Implementierung	**3**	10	30	10	30	10	30
Datengröße	**3**	10	30	10	30	10	30
Handling von Anhängen	**2**	5	10	9	18	7	14
Lizenzmodell	**2**	9	18	9	18	9	18
Berechtigungssystem	**2**	7	14	9	18	7	14
Gestaltungsmöglichkeiten	**1**	8	8	10	10	8	8
Einsetzbarkeit	**1**	10	10	10	10	10	10
Autorensystem	**1**	6	6	8	8	8	8
SUMME			460		477		472

Tabelle 10: Ergebnisse der Nutzwertanalyse

6.3 Fazit

Das Verteilen von Wissen spielt in Unternehmen eine sehr große Rolle, denn Wissen ist, wie sich herausgestellt hat, die wichtigste Ressource in einem Unternehmen. Verlässt ein Mitarbeiter das Unternehmen, muss sichergestellt sein, dass sein Wissen nicht verloren geht. Um diesen Schutz zu gewährleisten, muss das Wissen weitergegeben und im Unternehmen gespeichert werden.

Die Bewahrung des Wissens ist einer der sechs Kernprozesse des Wissensmanagements nach Probst, Raub und Romhardt. Um in einem Unternehmen Wissensmanagement effizient betreiben zu können und wichtiges Wissen, sowie Informationen zu bewahren und zu verteilen, eignet sich der Einsatz einer Dokumentationsplattform, auf welche die Mitarbeiter Zugriff haben. Bei einem Medium, welches solche essentiellen Informationen enthält, müssen natürlich einige Kriterien erfüllt sein, die diese wichtige Ressource schützen und die eine regelmäßige Aktualisierung und Erweiterung des Wissens ermöglichen.

Zur Definition von Kriterien wird auf die Theorie des Anforderungsmanagements zurückgegriffen. Durch den Einsatz von Satzschablonen bei der Erstellung eines Kriterienkatalogs, der als Analyseinstrument dient, können Missverständnisse vermieden werden und Kriterien für alle Stakeholder verständlich formuliert werden. Dementsprechend kann der Erfolg des Projektes besser gesteuert werden und die Anforderungen an eine Dokumentationsplattform können in einen strukturierten Rahmen eingebettet werden.

Dokumentationsplattformen zur Dokumentation und zum Verteilen von Wissen sind vielfältig am Markt vorhanden. Einige davon ähneln sich sehr, insbesondere die Wiki Plattformen haben viele Gemeinsamkeiten.

Um den Mitarbeitern der .Versicherung eine Alternative zu der bisher genutzten abteilungsinternen Notesdatenbank vorschlagen zu können, wurde eine Marktanalyse auf Basis des entwickelten Kriterienkatalogs durchgeführt. Eine ausführliche IST-SOLL-Analyse diente zur Formulierung wichtiger Anforderungen an das neue Dokumentationssystem, die im Kriterienkatalog festgehalten wurden. Im Rahmen dieser Marktanalyse wurden fünf Open Source Dokumentationsplattformen genauer betrachtet und analysiert.

Zu Beginn der Marktanalyse war das Angebot an Dokumentationsplattformen am Markt ziemlich unübersichtlich und beinahe zu umfangreich. Allerdings konnten einige Tools aussortiert werden, da sie den K.O.- Kriterien nicht entsprachen. Zuvor definierte K.O.-Kriterien sind zunächst einmal die Ablage von Dateien, eine einfach Navigation, sowie, dass das Produkt ein Open Source Produkt ist. Es hat sich herausgestellt, dass der Begriff „Open Source" für einige Unternehmen unter unterschiedlicher Definition ist. Während viele

Hersteller eine Open Source Plattform als kostenlos und quelloffen beschreiben, bezeichnen andere ein Tool als Open Source, wenn lediglich der Quellcode veröffentlich ist. Durch die Analyse verschiedener Dokumentationsplattformen ist jedoch deutlich geworden, dass es nicht unbedingt nötig ist, Geld für eine Dokumentationsplattform zu investieren, da der Markt genügend gute Dokumentationsplattformen bietet, die kostenlos sind.

Trotz der Schwierigkeit, dass die Tools (ohne einen Server) nicht heruntergeladen und dadurch auch nicht getestet werden konnten, hat die Marktanalyse gute Resultate ergeben. Es hat sich herausgestellt, dass insbesondere die Tools Alfresco, Agorum und DokuWiki sich besonders für die Zwecke der .Versicherung eignen, da sie allen Ansprüchen des Kriterienkataloges entsprachen. Zu beachten ist allerdings, dass Alfresco, im Gegensatz zu den anderen Dokumentationsplattformen kostenpflichtig ist. Dadurch gewährleistet das Unternehmen allerdings auch einen umfangreichen Support. Die Nutzung von Agorum ist zwar kostenlos, aber regelmäßige Updates kosten auch bei dieser Software einen kleinen Beitrag, sodass DokuWiki die einzige komplett kostenlose Lösung ist.

Die Tools MediaWiki und Wordpress hingegen entsprechen den Anforderungen nicht ausreichend, sodass sie ungeeignet für den Einsatz in der .Versicherung sind.
MediaWiki ist zwar ein sehr erfolgreiches Tool, allerdings wurde es hauptsächlich für sehr große Serverfarmen entwickelt und ist auf kleineren Servern nicht optimal zu implementieren. Außerdem bringt MediaWiki den Nachteil mit sich, dass kein Datenschutz gewährleistet ist, da Jeder die Inhalte lesen kann.
Wordpress eignet sich nicht für die Zwecke der .Versicherung , da die Software funktionsbedingt ein Blog ist und die Implementierung und die Strukturierung, die vor der Nutzung als Dokumentationsplattform anfällig wäre, sehr aufwendig und langwierig ist.

Die Entscheidung liegt nun zwischen Wiki Software (DokuWiki) und Dokumentenmanagementsystemen (Alfresco und Agorum) und zwischen kostenlos und kostenpflichtig. Im Rahmen einer Nutzwertanalyse hat sich herausgestellt, dass diese drei möglichen Dokumentationsplattformen in ihrer Bewertung sehr dicht beieinander liegen. DokuWiki führt dieses Ranking knapp an, gefolgt von Alfresco und Agorum. Alle drei Tools kommen für den Einsatz in der Anwendungsentwicklung der .Versicherung in Frage. In Anbetracht der Tatsachen, dass DokuWiki die einzige komplett kostenlose Lösung ist und Unternehmen, die Open Source Software einsetzen möchten, größtenteils aus der Motivation, Kosten zu reduzieren, handeln, ist es nach Bewertung der Autoren dieser Arbeit das Tool DokuWiki, welches am besten für die Zwecke der Anwendungsentwicklung der .Versicherung geeignet ist. Alfresco und Agorum können unter Berücksichtigung einer TCO-Analyse für den Einsatz bei der .Versicherung ebenfalls in Betracht gezogen werden.

Anhang

Anhangverzeichnis

Anhang 1: Dokumenttypen von Agorum[230]

von Format	in HTML 1)	in PDF	in TXT (für Suche)
ANS		x	x
BMP		x	x (mit OCR)
C		x	x
CPP		x	x
CSV			x
DOC	x	x	x
DOCM	x	x	x
DOCX	x	x	x
DOT	x	x	x
DOTM	x	x	x
DOTX	x	x	x
GIF		x	x (mit OCR)
H		x	x
HTM		x	x
HTML		x	x
I		x	x
JAVA		x	x
JPG		x	x (mit OCR)
JPEG			x (mit OCR)
JSP			x
LATEX			x
ODB			x
ODC			x
ODF		x	x
ODG	x	x	x
ODI	x	x	x
ODM	x	x	x
ODP	x	x	x
ODS	x	x	x
ODT	x	x	x
OTC			x
OTF		x	x
OTG	x	x	x
OTH			x
OTI	x	x	x
OTP	x	x	x
OTS	x	x	x
OTT	x	x	x
P		x	x
PCX			x (mit OCR)

230 Vgl. agorum® Software GmbH (o.J.c)

PDF	x	x	x
PNG			x (mit OCR)
POT	x	x	x
POTM	x	x	x
POTX	x	x	x
PPAM	x	x	x
PPS	x	x	x
PPT	x	x	x
PPTM	x	x	x
PPTX	x	x	x
PSD		x	x (mit OCR)
RTF		x	x
SGML			x
SQL			x
STC	x	x	x
STD	x	x	x
STI	x	x	x
STW	x	x	x
SXC	x	x	x
SXD	x	x	x
SXG	x	x	x
SXI	x	x	x
SXM		x	x
SXW	x	x	x
TEX			x
TGA			x (mit OCR)
TIF		x	x (mit OCR)
TIFF			x (mit OCR)
TXT		x	x
VRML			x
VSD			x
XLA	x	x	x
XLS	x	x	x
XLSB	x	x	x
XLSM	x	x	x
XLSX	x	x	x
XLT	x	x	x
XLTM	x	x	x
XLTX	x	x	x
XML		x	x

Quellenverzeichnisse

Literaturverzeichnis

Ahlert, M./Blaich, G./Spelsiek, J. (2006): Vernetztes Wissen. Organisationale, motivationale, kognitive und technologische Aspekte des Wissensmanagements in Unternehmensnetzwerken. 1. Aufl. Wiesbaden: Dt. Univ.-Verl. (Gabler-Edition Wissenschaft : Unternehmenskooperation und Netzwerkmanagement)

Al-Hawamdeh, S. (2003): Knowledge management. Cultivating knowledge professionals. Oxford: Chandos (Chandos information professional series)

Arndt, H. (2006): Integrierte Informationsarchitektur, 1. Aufl, Berlin: Springer

Blair, D. C. (2002): Knowledge management: Hype, hope, or help?: Wiley Subscription Services, Inc., A Wiley Company

Booch, G./ Rumbaugh, J., Jacobson, I. (2006): Das UML Benutzerhandbuch - Die unverzichtbare Referenz!: Aktuell zur Version 2.0 (Programmer's Choice). 1. Aufl. Addison-Wesley Verlag

Brem, geb. Krämer, S. (2009): Total Cost of Ownership als Instrument des Beschaffungsmanagements: Eine theoretisch-explorative empirische Untersuchung, Hamburg: diplom.de

Broßmann, M./ Mödinger, W. (2008): Praxisguide Wissensmanagement. Qualifizieren in Gegenwart Und Zukunft. Planung, Umsetzung Und Controlling in Unternehmen. (o.O.): Springer-Verlag New York Inc.

Der Brockhaus (1998): Leipzig [u.a.]: F. A. Brockhaus

Frey-Luxemburger, M. (2014): Wissensmanagement. - Grundlagen und praktische Anwendung. 2. Aufl. Wiesbaden: Springer Vieweg

Gehle, M. (2006): Internationales Wissensmanagement. Zur Steigerung der Flexibilität und Schlagkraft wissensintensiver Unternehmen. 1. Aufl. Wiesbaden: Dt. Univ.-Verl. (Gabler Edition Wissenschaft)

Gretsch, S. M. (2014): Wissensmanagement im Arbeitskontext: Bedarfsanalyse, Implementation eines Expertenfindungstools und Analyse zum Help-Seeking-Prozess. München: Springer Fachmedien Wiesbaden

Gust von Loh, S. (2009): Evidenzbasiertes Wissensmanagement. Wiesbaden: Betriebswirtschaftlicher Verlag Gabler

Hansch D./Schnurr H./Pissierssens P (2009): Semantic Media Wiki+ als Wissensplattform

Hasler Roumois, U. (2007): Studienbuch Wissensmanagement. Grundlagen der Wissensarbeit in Wirtschafts-, Non-Profit- und Public-Organisationen. Zürich: Orell Füssli (UTB, 2954)

Herrington, J. D.; Lang, J. W. (2006): PHP Hacks, 1. Aufl, Köln: O'Reilly Verlag

Hugos, M./Hulitzky, D. (2011): Business in the cloud. What every business needs to know about cloud computing. Hoboken, N.J: J. Wiley & Sons

Kenning, P./Schütte, R./ Blaich, G. (2003): Status Quo des Wissensmanagements im Dienstleistungssektor. Unter Mitarbeit von Ahlert, D./Zelewski, S. Nr. 3. Münster: MOTIWIDI Projektbericht

Kusterer, S. (2008): Qualitätssicherung im Wissensmanagement. Eine Fallstudienanalyse. 1. Aufl. Wiesbaden: Gabler (Gaber-Edition Wissenschaft)

Lucko, S./Trauner, B. (2005): Wissensmanagement: 7 Bausteine für die Umsetzung in der Praxis. München: Hanser

Marunde, G. (2003): Analyse von Methoden zur Suche in Portalplattformen und deren technische Integration am Beispiel der Portalplattform Up2gate.com, o. O.: diplom.de

Mehlan, A. (2007): Praxishilfen Controlling, o. O.: Haufe-Mediengruppe

Mujan, D. (2006): Informationsmanagement in lernenden Organisationen. Erzeugung von Informationsbedarf durch Informationsangebot: was Organisationen aus der Informationsbedarfsanalyse lernen können. Berlin: Logos-Verl.

Niebisch, T. (2013): Anforderungsmanagement in sieben Tagen, 1. Aufl. Berlin Heidelberg: Springer Gabler

Probst, G. J. B./ Raub, S./ Romhardt, K. (2012): Wissen managen. Wie Unternehmen ihre wertvollste Ressource optimal nutzen. 7., überarb. und erw. Aufl. Wiesbaden: Gabler

Probst, G. J. B./ Raub, S./ Romhardt, K. (2006): Wissen managen. Wie Unternehmen ihre wertvollste Ressource optimal nutzen. 5., überarb. Aufl. Wiesbaden: Gabler

Reiber, W. (2013): Vom Fachexperten zum Wissensunternehmer. Wissenspotenziale stärker nutzen, die persönliche Wirksamkeit erhöhen. Wiesbaden: Springer Fachmedien Wiesbaden; Imprint: Springer Gabler (SpringerLink : Bücher)

Reinmann, G./ Mandl, H. (Hg.) (2004): Psychologie des Wissensmanagements. Perspektiven, Theorien und Methoden. Göttingen: Hogrefe Verlag

Rupp C./ die SOPHISTen (2009): Requirements Engineering und –management. Aufl. 5. München/ Wien: Hanser

Rupp C./ Queins S. (2012): UML 2 glasklar: Praxiswissen für die UML-Modellierung. Aktualisierte und erweiterte Aufl. 4. München: Carl Hanser Verlag GmbH & Co. KG

Weber, J. (2007): Die Nutzwertanalyse zur Beurteilung von Entscheidungsalternativen im öffentlichen Sektor, o. O.: GRIN Verlag

Internetquellen

agorum® Software GmbH (Hrsg.) (2014): Agorum Broschüre, agorum ® core Open und Pro, Funktionen und Vorteile, http://www.agorum.com/fileadmin/img_agorum/pdf/agorum_broschuere_funktionen_vorteile_unterschiede.pdf, Abruf: 11.01.2015

agorum® Software GmbH (Hrsg.) (o.J.a): Agorum Website, Agorum Core Open Source, http://www.agorum.com/startseite/produkte/dms-ecm-agorum-core-open-source.html, Abruf: 27.12.2014

agorum® Software GmbH (Hrsg.) (o.J.b): Agorum Website, Funktionsübersicht, http://www.agorum.com/startseite/produkte/dms-ecm-agorum-core-open-source/funktionsuebersicht-agorum-core.html, Abruf: 27.12.2014

agorum® Software GmbH (Hrsg.) (o.J.c): Agorum Website, Unterstütze Dokumentenformate, http://www.agorum.com/startseite/produkte/dms-ecm-agorum-core-open-source/funktionsuebersicht-agorum-core/unterstuetzte-dokumentenformate-in-agorum-core.html, Abruf: 10.01.2015

agorum® Software GmbH (Hrsg.) (o.J.d): Agorum Website, Screenshots und Bilder, http://www.agorum.com/startseite/produkte/screenshots-bilder-zum-dms-ecm-agorum-core.html, Abruf: 11.01.2015

agorum® Software GmbH (Hrsg.) (o.J.e): Agorum Website, Infos zum Unternehmen, http://www.agorum.com/startseite/agorum-hautnah/infos-zum-unternehmen-agorum.html, Abruf: 11.01.2015

agorum® Software GmbH (Hrsg.) (o.J.f): Agorum Website, Referenzen, http://www.agorum.com/startseite/agorum-hautnah/kunden-referenzen-agorum-core.html, Abruf: 11.01.2015

Alfresco (Hrsg), (o.J): Open Source - Enterprise Content Management, http://www.ancud.de/Downloads/Alfresco.pdf, Abruf: 11.1.2015

Alfresco Software, Inc. (Hrsg.) (o.J): In die Cloud erweitertes Content Management, http://pages.alfresco.com/rs/alfresco/images/de-overview.pdf, Abruf: 12.1.2015

Alfresco Software, Inc. (Hrsg.) (2015): 100% Open Source, Continuous Innovation, http://www.alfresco.com/products/community, Abruf: 11.1.2015

Alfresco Software, Inc. (Hrsg.) (2015): Alfresco für das Dokumenten-Management im Finanzdienstsektor, https://www.alfresco.com/de/alfresco-fur-das-dokumenten-management-im-finanzdienstsektor, Abruf: 11.1.2015

Alfresco Software, Inc. (Hrsg.) (2015): Alfresco für das Dokumenten-Management im Finanzdienstsektor, https://www.alfresco.com/de/alfresco-fur-das-dokumenten-management-im-finanzdienstsektor, Abruf: 11.1.2015

Alfresco Software, Inc. (Hrsg.) (2015): Featured Resources, http://www.alfresco.com/, Abruf: 11.1.2015

Alfresco Software, Inc. (Hrsg.): Kunden-Support-Services, http://www.alfresco.com/de/technischer-support, Abruf: 11.1.2015

Alfresco Software, Inc. (Hrsg.), (2015): Using Alfresco, http://docs.alfresco.com/5.0/concepts/master-using-intro.html, Abruf: 11.1.2015

Ancud IT (Hrsg.), (o.J.): Alfresco, Open Source -Enterprise Content Management, http://www.ancud.de/Downloads/Alfresco.pdf, Abruf: 11.1.2015.

Apache Software Foundation (Hrsg.) (o.J): Apache Lucene Core, http://lucene.apache.org/core/, Abruf: 11.1.2015

Automattic (Hrsg.) (o.J.a): WordPress.com, Unterstütze Dateitypen, http://en.support.wordpress.com/accepted-filetypes/, Abruf: 10.01.2015

Automattic (Hrsg.) (o.J.b): WordPress.com, Benutzerrollen, http://de.support.wordpress.com/user-roles/, Abruf: 10.01.2015

Carl D./ Eidenberger H./Ludewig M./u.a. (2008): Open Source im Unternehmen, Mit heißer Feder, http://www.heise.de/open/artikel/MediaWiki-224566.html, Abruf: 10.1.2015

CosmoCode GmbH (Hrsg.) (2014): WikiMatrix / Statistics, http://www.wikimatrix.org/statistic/Most+Views, Abruf: 13.01.2015

DokuWiki (Hrsg.) (o. J.): DokuWiki Leaflet, https://www.dokuwiki.org/_media/leaflet-de.pdf, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2011a): DokuWiki Webseite - Ältere Versionen, https://www.dokuwiki.org/de:attic, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2011b): DokuWiki Webseite - Diff Funktion, https://www.dokuwiki.org/de:diff, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2012a): DokuWiki Webseite – Geschichte, https://www.dokuwiki.org/history_and_foss, Abruf: 11.01.2015

DokuWiki (Hrsg.) (2012b): DokuWiki Webseite - Systemvoraussetzungen und Anforderungen, https://www.dokuwiki.org/de:requirements, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2013a): DokuWiki Webseite - Access Keys, https://www.dokuwiki.org/de:accesskeys, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2013b): DokuWiki Webseite - Breadcrumbs-Navigation, https://www.dokuwiki.org/de:breadcrumbs, Abruf: 13.01.2015.

DokuWiki (Hrsg.) (2013c): DokuWiki Webseite - Letzte Änderungen, https://www.dokuwiki.org/de:recent_changes#unterschiede_zwischen_letzte_aenderungen_und_aeltere_versionen, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2013d): DokuWiki Webseite - Login und Registrierung, https://www.dokuwiki.org/de:login, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2013e): DokuWiki Webseite – Plugins, https://www.dokuwiki.org/de:plugins, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2013f): DokuWiki Webseite – Zugriffskontrolle, https://www.dokuwiki.org/de:acl, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2014a): DokuWiki Webseite, https://www.dokuwiki.org/dokuwiki, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2014b): DokuWiki Webseite – Installation, https://www.dokuwiki.org/start?id=de:install, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2014c): DokuWiki Webseite - Installation von Templates, https://www.dokuwiki.org/de:template, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2014d): DokuWiki Webseite - Letzte Änderungen, https://www.dokuwiki.org/de:toolbar?do=recent, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2014e): DokuWiki Webseite - Open Source, https://www.dokuwiki.org/start?id=de:donate, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2014f): DokuWiki Webseite – Releases, https://www.dokuwiki.org/start?id=de:changes, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2014g): DokuWiki Webseite – Toolbar, https://www.dokuwiki.org/de:toolbar, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2014h): DokuWiki Webseite - Umgang mit Bildern und Medien, https://www.dokuwiki.org/de:images, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2015a): DokuWiki Webseite - Eigenschaften und Funktionen, https://www.dokuwiki.org/start?id=de:features, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2015b): DokuWiki Webseite - Medien Manager, https://www.dokuwiki.org/features?do=media&ns=, Abruf: 13.01.2015

DokuWiki (Hrsg.) (2015c): DokuWiki Webseite – Namensräume, https://www.dokuwiki.org/de:namespaces, Abruf: 13.01.2015

Humpa, M. (o. J.): Scribble Papers, http://www.chip.de/downloads/Scribble-Papers_28455266.html, Abruf: 16.01.2015

Inpsyde GmbH (Hrsg.) (2013): WordPress 10th Anniversary, https://marketpress.de/files/2013/05/wordpress_10_years_anniversary_infographic_by_marketpresscom.jpg, Abruf: 11.01.2015

Inpsyde GmbH (Hrsg.) (o.J.a): WordPress Deutschland, Funktionen, http://wpde.org/funktionen/, Abruf: 10.01.2015

Inpsyde GmbH (Hrsg.) (o.J.b): WordPress Deutschland, Open Source, http://wpde.org/open-source/, Abruf: 10.01.2015

Inpsyde GmbH (Hrsg.) (o.J.c): WordPress Deutschlad, Voraussetzungen, http://wpde.org/voraussetzungen, Abruf: 11.01.2015

Inpsyde GmbH (Hrsg.) (o.J.d): WordPress Deutschland, Maximale Uploadgröße, http://forum.wpde.org/konfiguration/85610-maximale-uploadgroesse-aendern.html, Abruf: 11.01.2015

Inpsyde GmbH (Hrsg.) (o.J. e): WordPress Deutschland, Designanpassungen und Themes, http://wpde.org/themes/, Abruf: 11.01.2015

Inpsyde GmbH (Hrsg.) (o.J.f): WordPress Deutschland, Mobil mit WordPress, http://wpde.org/mobil/, Abruf: 11.01.2015

Inpsyde GmbH (Hrsg.) (o.J.g): WordPress Deutschland, Plugins, https://wordpress.org/plugins/, Abruf: 12.01.2015

Media Wiki (Hrsg.) (2014): Manual: Configuring file uploads/de, http://www.mediawiki.org/wiki/Manual:Configuring_file_uploads/de, Abruf: 10.1.2015

Media Wiki (Hrsg.) (2014): Help: Searching/de, http://www.mediawiki.org/wiki/Help:Searching/de, Abruf: 11.1.2015

Media Wiki (Hrsg.) (2014): Manual: $wgFileBlacklist, http://www.mediawiki.org/wiki/Manual:$wgFileBlacklist, Abruf: 10.1.2015

Media Wiki (Hrsg.) (2014): Manual: Image administration/de, http://www.mediawiki.org/wiki/Manual:Image_administration/de, Abruf: 11.1.2015

Media Wiki (Hrsg.) (2014): Handbuch: Was ist Media Wiki?, http://www.mediawiki.org/wiki/Manual:What_is_MediaWiki%3F/de, Abruf: 11.01.2015

Media Wiki (Hrsg.) (2014): Help: Managing files, http://www.mediawiki.org/wiki/Help:Managing_files, Abruf: 11.1.2015

Media Wiki (Hrsg.) (2014): Help: Navigation, http://www.mediawiki.org/wiki/Help:Navigation/de, Abruf: 11.1.2015

Merschmann H, (2008): Semantic Web: Das Internet soll klüger werden, http://www.spiegel.de/netzwelt/web/semantic-web-das-internet-soll-klueger-werden-a-561831.html, Abruf: 11.1.2015

o. V. (o. J.): WORDPRESS, https://de.wordpress.org/, Abruf: 16.01.2015

o. V. (2013): Das neue Dokumentationstool Udocs ist online, http://www.unicon-software.com/news/das-neue-dokumentationstool-udocs-ist-online/, Abruf: 16.01.2015

o. V. (2014b): Was ist Scribble Papers?, http://home.arcor.de/jhoetger/scpapers/, Abruf: 16.01.2015

o.V. (2014c): Manual: What is MediaWiki?, https://www.mediawiki.org/wiki/MediaWiki, Abruf: 16.01.2015

o.V. (2014d): Herzlich Willkommen bei Udocs – Die Dokumentation zu den Cloud-Computing-Lösungen von Unicon Software, http://www.unicon-software.com/udocs/de/, Abruf: 16.01.2015

o. V. (2015a): Alfresco Produkte, http://www.alfresco.com/de/node/3179, Abruf: 16.01.2015

perun.net webwork gmbh (Hrsg.) (2014): WordPress Bücher und Ebooks, Wie stellt man die ältere Version eines WordPress-Artikels wieder her?, http://www.wpbuch.de/2011/01/wie-stellt-man-die-aeltere-version-eines-wordpress-artikels-wieder-her/, Abruf: 11.01.2015

Peters, M. (2014): Eigenes Wiki erstellen: Die 3 besten Anbieter – CHIP, http://praxistipps.chip.de/eigenes-wiki-erstellen-die-3-besten-anbieter_36727, Abruf: 11.01.2015

Schwaninger, M. (2000): Implizites Wissen und Managementlehre. organisationskybernetische Sicht. St. Gallen: Inst. für Betriebswirtschaft, http://www.ifb.unisg.ch/org/IfB/ifbweb.nsf/SysWebRessources/beitrag41/$FILE/DB41_ImplizitesWissen_def.pdf, Abruf: 04.01.2015

Schwerthalter, R. (Hrsg.) (2014): pressengers, So verbessert ihr die WordPress-Suche, http://pressengers.de/plugins/so-verbessert-ihr-die-wordpress-suche/, Abruf: 10.01.2015

Semantic Media Wiki (Hrsg.) (2012): Help: Bearbeiten, https://semantic-mediawiki.org/wiki/Help:Bearbeiten, Abruf: 11.1.2015

Semantic Media Wiki (Hrsg.) (2012): Help: Semantische Suche, http://semantic-mediawiki.org/wiki/Help:Semantische_Suche, Abruf: 11.1.2015

Thommen, J.-P. (2015): Gabler Wirtschaftslexikon, Stichwort Anspruchsgruppen, http://wirtschaftslexikon.gabler.de/Archiv/1202/anspruchsgruppen-v6.html, Abruf: 20.01.2015

Unitversität Kassel – IT Servicezentrum (Hrsg.) (o.J): Rollen in Alfresco, http://www.uni-kassel.de/its-handbuch/kommunikation/dms/rollen-in-alfresco.html, Abruf: 11.1.2015

Wikibooks (Hrsg.) (2013): Media Wiki/Vertiefung/Benutzerrechte, http://de.wikibooks.org/wiki/MediaWiki/_Vertiefung/_Benutzerrechte, Abruf: 11.1.2015

Wikipedia (Hrsg.) (2014): Datei:Alfresco-logo.svg, http://de.wikipedia.org/wiki/Datei:Alfresco-logo.svg, Abruf: 19.1.2015

Wikipedia (Hrsg.) (2014): Media Wiki, http://de.wikipedia.org/wiki/MediaWiki#cite_note-3, Abruf: 25.1.2015

Wikipedia (Hrsg.) (2015): Wiki, http://de.wikipedia.org/wiki/Wiki, Abruf: 11.1.2015

Werner B. (2009): Speicherort der Dateien, https://forums.alfresco.com/de/speicherort-der-dateien-09032009-1435, Abruf: 19.1.2015

Gesprächsverzeichnis

Hötger, J. (2015): Erfinder und Entwickler von Scribble Papers, E-Mail Gespräch am 11.01.2015

Anonymisiert, U. (2014): Repräsentant der Anwendungsentwicklung der .Versicherung , Suttgart, persönliches Gespräch am 12.12.2014

Anonymisiert, U. (2015): Repräsentant der Anwendungsentwicklung der .Versicherung , E-Mail Gespräch vom 04.01.2015 – 19.01.2015

Henke, A. (2015): Administratorin des Forums und WikiManagerin von DokuWiki, Schriftliche Beantwortung von Fragen, 10.01.2015

Pauka J. (2015): Alfresco, Schriftliche Beantwortung von Fragen, 6.1.2015

Schulze, O. (2014): Mitarbeiter von Agorum, Email Gespräch am 28.12.2014

Einsatz von Open Source Tools zur PDF-Erzeugung bei Versicherungen

Integrationsprojekt

vorgelegt am 04.02.2015

Fakultät Wirtschaft

Studiengang Wirtschaftsinformatik – Application Management

Kurs WWI2012E

von

Patrick Hartmann, Sammy Kessira, Franziska Losch und Daniel Paukner

Partnerunternehmen: DHBW Stuttgart:

.Versicherung

Inhaltsverzeichnis

Abkürzungsverzeichnis

AGB	= Allgemeine Geschäftsbedingungen
BITKOM	= Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V.
BSD	= Berkeley Software Distribution
DHBW	= Duale Hochschule Baden-Württemberg
GPL	= GNU General Public License
KOS	= Kompetenzzentrum Open Source
LGPL	= Lesser General Public License
MS	= Microsoft
OCR	= Optical Character Recognition
OSS	= Open Source Software
OSI	= Open Source Initiative
Open LDAP	= Open Lightweight Directory Access Point
PDF	= Portable Document Format
USP	= Unique Selling Proposition
SLA	= Service Level Agreement

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Problemstellung

Viele Versicherungsunternehmen setzen heutzutage getrieben von historischen Gründen verschiedenste Anwendungen zur unterschiedlichsten Erzeugung von PDF-Dokumenten ein. Dies reicht von der einfachen PDF-Erstellung eines Textdokuments an einem Einzelarbeitsplatz bis hin zur automatisierten Verarbeitung von archivierungskonformen Dokumenten. Weitestgehend weist die eingesetzte Software den passenden Funktionsumfang für einen Anwendungsbereich vor. Des Weiteren ist die Software überwiegend proprietär und im Hinblick auf die Anzahl an verschiedenen Produkten samt benötigter Lizenzierung sehr kostenintensiv. Aus diesem Grund stellt sich aus wirtschaftlicher Sicht die Frage, ob anstelle kommerzieller Software OS-Produkte eingesetzt werden können und dies bei gleichbleibender oder höherer Qualität.

1.2 Ziel der Marktanalyse

Das Ziel des Integrationsseminars ist die Gegenüberstellung eines funktionalen Überblicks von alternativen OS-Produkten mit der momentan eingesetzten proprietären Software seitens des Versicherungsunternehmens. Die durchgeführte Analyse soll für das Versicherungsunternehmen als Bewertungsgrundlage für eine Ablösung der kommerziellen Software und einer Implementierung von OS-Produkten dienen. Abschließend wird eine Handlungsempfehlung basierend auf den Unternehmensanforderungen ausgesprochen.

1.3 Vorgehensweise

Das Integrationsseminar wird in Form eines Projektes gehandhabt und somit in mehrere Phasen aufgeteilt. In der Planungsphase werden im Dialog mit dem Versicherungsunternehmen die speziellen Anforderungen und Merkmale an die OS-Produkte definiert. Im gleichen Zuge wird ein Projektplan erstellt, welcher zur Fortschrittsverfolgung für das Versicherungsunternehmen und dem Projektteam dient. Anschließend werden die zuvor besprochenen Kriterien ausgearbeitet, gewichtet und in die Form eines Kriterienkatalogs, womit die Testprodukte beurteilt werden, gebracht. Der Inhalt des Katalogs besteht aus fundamentalen und unternehmensspezifischen Anforderungen. Gleichzeitig startet die Recherche nach OS-Software zur PDF-Erzeugung. In der Durchführungsphase werden die OS-Produkte analysiert und bewertet. Gleiches gilt für das im Versicherungsunternehmen eingesetzte Tool. Folglich werden die getesteten Produkte gegenübergestellt, anhand jener erkannt werden kann, ob OS-Software dem kommerziellen Produkt ebenwürdig oder gar zu bevorzugen ist. Abschließend wird dem Versicherungsunternehmen eine Handlungsempfehlung ausgesprochen.

1.4 Projektarbeitsumgebung

Das Projekt wurde vom Kompetenzzentrum Open Source (KOS) der Dualen Hochschule Baden-Württemberg (DHBW) ins Leben gerufen. Unter dem Forschungsauftrag "Open Source" gibt es mehrere verschiedene Fragestellungen, welche zu Integrationsseminaren mit dem jeweiligen Versicherungsunternehmen und DHBW-Studenten führt. Das Ziel der Projekte ist die Untersuchung und Identifizierung der Einsatzfelder für OS-Software in Unternehmen. Hierbei liegt das Hauptaugenmerk auf der Reduzierung der Lizenzkosten unter der Berücksichtigung von unternehmensspezifischen Anforderungen.

2 Theoretische Abhandlung

2.1 Marktanalyse

In einer Marktanalyse werden verschiedenste Informationen gesammelt, um durch weitere Auswertung einen genauen Überblick zu einem bestimmten Sachverhalt zu erlangen. So kann diese Methode bei einer Produktinnovation, einer Unternehmensanalyse oder – wie in diesem Fall – einer Produktauswahl Anwendung finden. In jedem Fall gilt, dass der Markt, seine Anbieter und deren Produkte bekannt sind und fundierte Informationen zur Verfügung stehen. Im Fall einer Einführung eines neuen Produktes ist es daher unabdingbar, dass der Markt auf Konkurrenzprodukte geprüft wird und deren Qualität und Eigenschaften verglichen werden, um so letztendlich die Vor- und Nachteile des eigenen Produktes oder gar einen Unique Selling Proposition (USP) zur Differenzierung herauszufinden.[1] Besonders durch die Globalisierung und der daher gehenden Marktsättigung gewann diese Thematik an Wichtigkeit, um den Kunden bzw. Anwendern die Vorzüge des eigenen Produktes näherzubringen. In derselben Weise wird bei der Auswahl von Software für ein Unternehmen vorgegangen. Durch das Internet gibt es ein fast unendliches Angebot an verschiedensten Anwendungen, welche sich nicht alle gleichermaßen für einen Unternehmenseinsatz eignen. Hierbei gilt das optimale Produkt zu eruieren, welches zu den Unternehmensanforderungen passt. Zuvor muss die marktrelevante Software anhand grundliegender Vorgaben recherchiert werden, wie beispielsweise darf keine proprietäre Software gewählt werden. Im Folgenden wird näher auf die Teile einer Marktanalyse eingegangen.[2]

2.1.1 Prozess einer Marktanalyse

Eine Marktanalyse lässt sich in sechs Prozessschritten beschreiben. Zuerst muss das vorliegende Problem bewusst wahrgenommen und anschließend beschrieben werden. Daraufhin werden entsprechende Alternativen zur Lösung der Problemstellung gesucht und anschließend im Kontext der Anforderungen bewertet. Anhand der vorgenommenen Beurteilung wird eine Handlungsempfehlung ausgesprochen bzw. eine Entscheidung für eine der analysierten Lösungsalternativen gewählt. Im nächsten Schritt wird die gewählte Lösung in die Praxis umgesetzt, wodurch das Problem bei einer ordentlich durchgeführten Marktanalyse behoben ist. Je nach Problemstellung kann der Erfolg der Lösung in Form einer Soll-Ist-Analyse beliebig oft gemessen werden, um gegebenenfalls Nachbesserungspotential zu erkennen. Jeder der genannten Prozessschritte setzt in aller Regel viele zur Verfügung stehende Informationen voraus. Nachfolgend sind die sechs Prozessschritte aufgelistet:

1. Identifikation und Beschreibung des Problems

[1] Grimm, R. / Schuller, M. / Wilhelmer, R. (2014), S. 97 ff.
[2] Vgl. Berekoven, L. / Eckert, W. / Ellenrieder, P. (2009), S. 334 ff.

2. Planung von Lösungsmöglichkeiten
3. Beurteilung der Lösungsmöglichkeiten
4. Entscheidung für eine Lösungsmöglichkeit
5. Implementierung der Lösung
6. Überprüfung des Erfolgs[3]

2.1.2 Bewertung von Informationen

Um eine repräsentative Marktanalyse zu erhalten, müssen die verfügbaren Informationen zur Entscheidungsfindung geeignet sein. In der Regel erleichtert und verbessert die Qualität der Informationen die Entscheidung für eine Lösung. Hierzu müssen die Informationen einem Bewertungsmaßstab unterzogen werden, wobei davon zwei Kriterien Priorität haben. Zum einen die Informationsqualität in Bezug auf objektive Messkriterien und die Relevanz zur Problemstellung. Zum anderen muss der ökonomische Aspekt betrachtet werden, denn die Recherche von Informationen für eine Marktforschung kann kostspielig sein und daher muss die Erbringung einer Information unter Abwägung einer Kosten-Nutzen-Relation resultieren. Bei der Bewertung der Informationen können folgende qualitative Bewertungskriterien beachtet werden:

- Nützlichkeit
- Vollständigkeit
- Aktualität
- Wahrheit

Betriebswirtschaftlich betrachtet sind Informationen zu beschaffen, wenn die Kosten geringer als die Erträge der Verwendung sind. Als Beispiel kann eine Information herangezogen werden, welche einen Auftrag bei einem Großkunden sichert. Hierbei können leicht die Kosten zur Informationsbeschaffung von der Ertrag des gewonnenen Kundenauftrags abgezogen werden und sofern summa summarum der Betrag positiv ist wäre die Information rentabel.[4]

2.1.3 Skalierungsverfahren und Produkttest

Das Skalierungsverfahren bezieht sich auf theoretische, nicht beobachtbare Sachverhalte. Diese Sachverhalte werden innerhalb einer Person wirksam und werden auch hypothetische Konstrukte oder intervenierende Variable wie Emotionen, Einstellung oder Präferenzen ge-

[3] Vgl. Berekoven, L. / Eckert, W. / Ellenrieder, P. (2009), S. 19 ff.
[4] Vgl. Berekoven, L. / Eckert, W. / Ellenrieder, P. (2009), S. 22 ff.

nannt. Aus diesem Grund werden diese qualitativen Merkmale in quantitative Werte skaliert. Folgend sind die gebräuchlichen Skalierungsverfahren im Überblick:

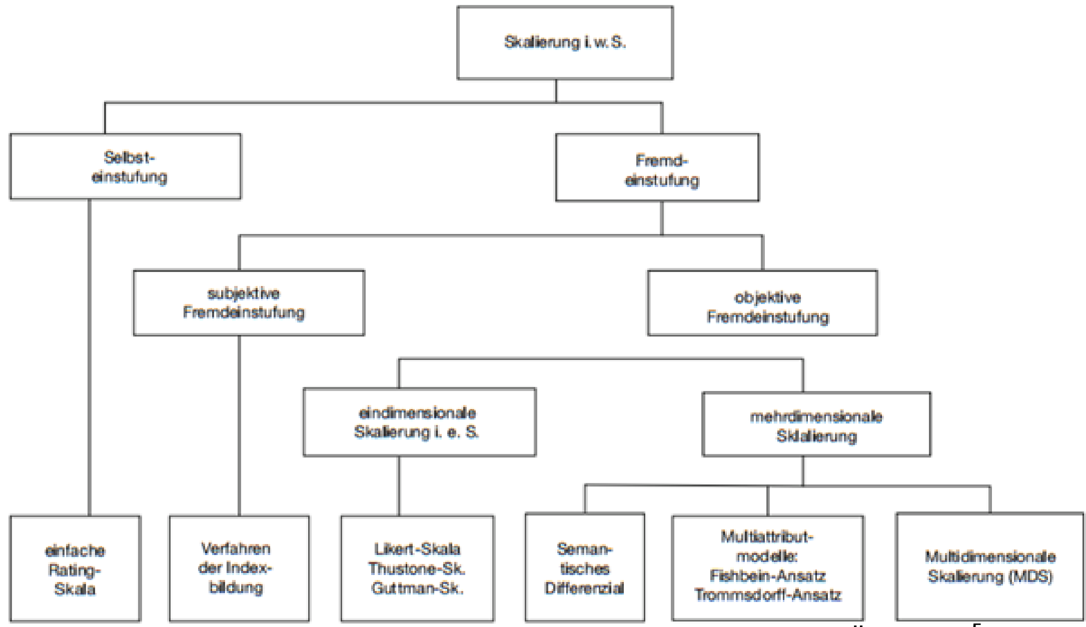


Abbildung 1: Die gebräuchlichen Skalierungsverfahren im Überblick[5]

In dieser Arbeit für das Integrationsseminar wird jedoch nur auf das Verfahren der Indexbildung eingegangen, da die Handhabung des Bewertungskatalogs daran angelehnt ist. Es charakterisiert sich durch die Freiheit für den Untersuchungsleiters, indem dieser frei wählen kann, was während des Verfahrens abgefragt wird. Zusätzlich liegt die Bestimmung der Gewichtung ebenso in seinem subjektiven Ermessen.

Ein Produkttest ist eine experimentelle Untersuchung eines existierenden Produkts oder Prototyps, bei der nach bestimmten Kriterien überprüft werden. Hierzu wird ein Konzepttest angefertigt wie ein Kriterienkatalog. Generelle Zielsetzungen eines Produkttests sind:

- Ermittlung von Produktalternativen
- Ermittlung des besten Produktes aus den Alternativen
- Überprüfung auf Erfüllung der Anforderungen[6]

2.1.4 Auswertung der erhobenen Daten

Die erhobenen Daten werden mittels eines Kriterienkatalogs bewertet. Anschließend kann ein direkter Vergleich der Testprodukte gezogen werden, wodurch das passendste Produkt ermittelt wird. Mit einer entsprechenden Handlungsempfehlung bzw. Implementierung wird die Marktanalyse abgeschlossen.

[5] Berekoven, L. / Eckert, W. / Ellenrieder, P. (2009), S. 67
[6] Berekoven, L. / Eckert, W. / Ellenrieder, P. (2009), S. 66

2.2 Kriterienkatalog

Ein Kriterienkatalog ist ein Tool zur Entscheidungsfindung. Es wird genutzt um zum Beispiel mehrere Programme und Tools miteinander zu vergleichen.

Kriterienkataloge stehen viel in der Kritik. Eine praktische Anwendung scheitert oft an abstrakten Formulierungen und daraus entstehender Unklarheit.[7] Richtig angewendet zeigt sich ein Kriterienkatalog jedoch als praktische Methode einer Entscheidungsfindung, zum Beispiel zur Auswahl einer möglichst geeigneten Software zur PDF-Erzeugung und Bearbeitung für ein Versicherungsunternehmen.

2.2.1 Erstellung eines Kriterienkatalogs

Um einen Kriterienkatalog zu erstellen, braucht es zunächst möglichst genau definierte Kriterien. Diese stammen meist aus einem Lasten- beziehungsweise Pflichtenheft.[8] Ein Lastenheft ist ein Dokument, indem Anforderungen genau beschrieben werden. Im Pflichtenheft werden dann die gewünschten Wege zur Umsetzung der Anforderungen beschrieben. Die aus dem Lasten- beziehungsweise Pflichtenheft entnommenen Kriterien sollten dann der Übersichtshalber in Hauptgruppen zusammengefasst, gewichtet und priorisiert werden. Im Beispiel der Entscheidungsfindung einer Software wäre eine Unterteilung der Kriterien in funktionale und nicht-funktionale Anforderungen sinnvoll. Die funktionalen Anforderungen bieten die beste Möglichkeit, Software miteinander zu vergleichen, aus den nicht-funktionalen Anforderungen kann unter anderem der Reifegrad der Software, der Hersteller und die Kosten extrahiert werden. Die Gewichtung und Priorisierung der Kriterien, in der unter anderem auch KO-Kriterien, also Kriterien, die wenn sie nicht erfüllt werden einen Ausschluss eines Produkts aus dem Vergleich bedeuten, ausfindig und gekennzeichnet werden sollten, unterstützt eine genaue Bewertung eines Produktes. Ebenso unterstützen viele aussagekräftige Kriterien eine Entscheidungsfindung.

Jeder Kriterienkatalog sieht anders aus, da sie immer an die vorhandenen Kriterien, Produkte, Präferenzen des Nutzers und so weiter angepasst werden. Tabelle 1 zeigt ein Beispiel für einen kleinen Kriterienkatalog:

[7] Vgl. Röder, H. / u.a. (o.J.)
[8] Vgl. bund.de (o.J.)

Kriterien	KO-Kriterium	Gewichtung	Bewertung
Hauptgruppe 1 (z.B. funktionelle Kriterien)		80 %	
Kriterium 1	X	35%	
Kriterium 2	X	35%	
Kriterium 3		10%	
Hauptgruppe 2 (z.B. nicht-funktionelle Kriterien)		20%	
Kriterium 4	X	15%	
Kriterium 5		5%	
Gesamtbewertung		100%	

Tabelle 1: Beispiel eines Kriterienkataloges

2.2.2 Vor- und Nachteile von Kriterienkatalogen

Kriterienkataloge bieten sowohl Vorteile, als auch Nachteile gegenüber anderen Verfahren zur Entscheidungsfindung. Ein großer Vorteil ist, dass die Erstellung und Bewertung entlang eines Kriterienkatalogs nur niedrige Kosten verursacht, da nur eine (fachlich qualifizierte) Arbeitskraft, eine Programmkopie, für einen Test der Software, und Hardware benötigt wird.[9] Der Kriterienkatalog ermöglicht einen übersichtlichen und sehr schnellen Vergleich mehrerer Produkte. Weitere Vorteile sind die Objektivität des Verfahrens sowie die Eingrenzung der Produktbreite durch Ausschluss- beziehungsweise KO-Kriterien.[10] Ebenso ist ein Kriterienkatalog sehr flexibel, da man je nach Bedarf Kriterien austauschen und die Gewichtung von Kriterien verändern kann. So kann zum Beispiel ein Kriterienkatalog, an dem nur die Gewichtung geändert wird, für verschiedene Abteilungen in einem Unternehmen genutzt werden.

Einer der größten Nachteile eines Kriterienkatalogs ist die mögliche Unvollständigkeit[11], das heißt, das Fehlen von eventuell wichtigen Kriterien. Ebenso ist die Subjektivität, geschuldet durch die Bearbeitung durch eine Person, ein Problem. Durch die Subjektivität besteht das Risiko, das Prioritäten und die Gewichtung falsch gesetzt wird, so dass zum Beispiel ein eigentlich unwichtiges Kriterium schwerer wiegt als ein Kriterium, das auch ein Ausschlusskriterium sein kann.

[9] Vgl. Jellito, M. (2002)
[10] Vgl. Weber, M. (2008)
[11] Vgl. Weber, M. (2008)

2.2.3 Fazit des Kriterienkatalogs

Zusammenfassend ist zu sagen, dass ein Kriterienkatalog ein einfacher und guter Weg ist um eine Entscheidungsfindung zu unterstützen. Es wird Objektivität, eine Eingrenzung an Produkten und niedriger monetärer als auch zeitlicher Aufwand geboten und die Nachteile können durch klar definierte, aussagekräftige und detailliert aufgeschriebenen Kriterien, die aus Lasten- und Pflichtenhefte abgeleitet werden, abgefangen werden. Ein Kriterienkatalog ist somit eine einfache, objektive Methode um verschiedene Produkte schnell zu vergleichen.

2.3 Open Source und Lizenzen

2.3.1 Definition / Was ist Open-Source

Open Source oder auch quelloffene Software ist eine der Erscheinungsformen sogenannter freier Software. Dabei liegt der Quelltext der Open Source Software (OSS) offen und steht unter einer Lizenz, die von der Open Source Initiative (OSI) anerkannt wird. Damit ist die Verwertung, Vervielfältigung und Bearbeitung des Quelltextes jedoch nicht vorbehaltslos gestattet, denn bei der Open Source Software wird vielfach die Einräumung von Nutzungsrechten von bestimmten Voraussetzungen abhängig gemacht.

Open Source Software lässt sich somit von anderen Softwaretypen wie Public Domain Software, Freeware und Shareware abgrenzen. Public Domain Software gibt dem Nutzer die Möglichkeit der uneingeschränkten und vorbehaltslosen Vervielfältigung, Verbreitung und Veränderung des Quelltextes. Bei Shareware unterliegt die Nutzung gewissen Beschränkungen wie z.B. in zeitlicher Hinsicht oder im Bezug auf die kommerzielle Verwertung. Im Vergleich dazu wird bei kostenlos vertriebener Freeware der Sourcecode nicht offen gelegt und es besteht keine Befugnis zur Änderung der Software.[12]

Die bereits erwähnte Open Source Initiative ist ein gemeinnütziger Zusammenschluss, dessen Ziel die Festlegung von einheitlichen Definitionen und eines einheitlichen Standards von Open Source Software ist. Die Open Source Definition der Open Source Initiative besteht aus 10 Kriterien, die im Folgenden beschrieben werden sollen.[13]

1. Freie Weitergabe

Die Lizenz darf niemanden darin hindern, die Software zu verkaufen oder sie mit anderer Software zusammen in einer Software-Distribution weiterzugeben. Die Lizenz darf außerdem für den Fall eines solchen Verkaufs keine Lizenz- oder sonstige Gebühren festschreiben.

[12] Vgl. BITKOM (2006), S.6
[13] Vgl. Open Source Initiative (2015)

2. Das Programm muss den Quellcode beinhalten

Die Weitergabe des Programms muss sowohl für den Quellcode, als auch für die kompilierte Form zulässig sein. Wenn das Programm in irgendeiner Form ohne Quellcode weitergegeben wird, so muss es eine allgemein bekannte Möglichkeit geben, den Quellcode zum Selbstkostenpreis zu bekommen. Dies kann zum Beispiel durch einen gebührenfreien Download aus dem Internet erfüllt werden. Der Quellcode soll die Form eines Programms sein, die ein Programmierer vorzugsweise bearbeitet. Absichtlich unverständlich geschriebener Quellcode ist daher nicht zulässig. Zwischenformen des Codes, so wie sie etwa ein Präprozessor oder ein Konverter erzeugt, sind unzulässig.

3. Abgeleitete Software

Die Lizenz muss Veränderungen und Derivate zulassen. Außerdem muss sie es zulassen, dass die solcherart entstandenen Programme unter denselben Lizenzbedingungen weiter vertrieben werden können wie die Ausgangssoftware.

4. Unversehrtheit des Quellcodes des Autors

Die Lizenz darf die Möglichkeit, den Quellcode in veränderter Form weiterzugeben, nur dann einschränken, wenn sie vorsieht, dass zusammen mit dem Quellcode sogenannte Patch Files weitergegeben werden dürfen, die den Programmcode bei der Kompilierung verändern. Die Lizenz muss die Weitergabe von Software, die aus verändertem Quellcode entstanden ist, ausdrücklich erlauben. Die Lizenz kann verlangen, dass die abgeleiteten Programme einen anderen Namen oder eine andere Versionsnummer als die Ausgangssoftware tragen

5. Keine Diskriminierung von einzelnen Personen oder Gruppen

Die Lizenz darf keinerlei Personen oder Gruppen diskriminieren und damit niemanden benachteiligen.

6. Keine Einschränkungen für bestimmte Anwendungsbereiche

Die Lizenz darf niemanden daran hindern, das Programm in einem bestimmten Bereich einzusetzen. Beispielsweise darf sie den Einsatz des Programms in einem Geschäft oder in der Genforschung nicht ausschließen.

7. Weitergabe der Lizenz

Die Rechte an einem Programm müssen auf alle Personen übergehen, die diese Software erhalten, ohne dass für diese die Notwendigkeit bestünde, eine eigene, zusätzliche Lizenz zu erwerben.

8. Die Lizenz darf nicht für ein bestimmtes Produkt gelten

Die Rechte an dem Programm dürfen nicht davon abhängig sein, ob das Programm Teil eines bestimmten Software-Paketes ist. Wenn das Programm aus dem Paket herausgenommen und im Rahmen der zu diesem Programm gehörenden Lizenz benutzt oder weitergegeben wird, so sollen alle Personen, die dieses Programm dann erhalten, alle Rechte daran haben, die auch in Verbindung mit dem ursprünglichen Software-Paket gewährt wurden.

9. Die Lizenz darf die Weitergabe nicht einschränken

Die Lizenz darf keine Einschränkungen enthalten bezüglich anderer Software, die zusammen mit der lizenzierten Software weitergegeben wird. So darf die Lizenz z.B. nicht verlangen, dass alle anderen Programme, die auf dem gleichen Medium weitergegeben werden, auch quelloffen sein müssen.

10. Die Lizenz muss technologie-neutral sein

Die Lizenz darf nicht auf eine individuelle Technologie oder einen bestimmten Interfacetyp beziehen.

2.3.2 Lizenzarten

Im Verlauf der Entwicklung von Open Source Software wurden verschiedene Lizenzbedingungen mit zum Teil unterschiedlichen Anforderungen an den Anwender entwickelt. Diese Lizenzbedingungen müssen mit den Kriterien der Open Source Definition der Open Source Initiative übereinstimmen, um als “Open Source Licenses” zu gelten. Damit die Lizenz offiziell als Open Source Lizenz akzeptiert wird, muss ein Überprüfungsprozess der Open Source Initiative stattfinden, welcher absichert, dass die Lizenz den existierenden Normen und Erwartungen entspricht.

Im Folgenden sollen einige der populärsten und am häufigsten verwendeten Lizenzarten genauer beschrieben werden.

2.3.2.1 „Copyleft“-Lizenzen (GNU General Public License, GPL)

Die GNU General Public License ist die am häufigsten verwendete Open Source Software-Lizenz, welche genutzt werden kann, um Software kostenlos zu nutzen, ändern und zu verbreiten. Die GNU GPL wurde 1989 von Richard Stallman, dem Gründer des GNU-Projektes geschrieben. Sie war damit der Prototyp für Lizenzen mit Copyleft-Effekt.[14]

[14] Vgl. Jaeger, T. / Schulz, C. (2005), S.29

Lizenzen mit Copyleft-Effekt verpflichten den Nutzer, die Weiterentwicklung der Open Source Software wiederum den Bestimmungen der Lizenz zu unterstellen, wenn ein unter der Lizenz stehender Code verändert wurde und daraus ein „abgeleitetes Werk" entsteht, welches weitergegeben oder vertrieben werden soll. Eine Verknüpfung der Open Source Software mit proprietärer Software, d.h. Software für deren Nutzung eine Lizenzgebühr erhoben wird, ist damit grundsätzlich nicht möglich. Wird die proprietäre Software direkt mit der Open Source Software verknüpft, wird diese ebenfalls eine Open Source Software. Damit soll verhindert werden, dass ein geändertes oder weiterentwickeltes Programm in proprietäre Software umgewandelt wird und daraufhin kommerziell vertrieben werden kann. Diesen Effekt der Lizenzbedingung bezeichnet man als „Copyleft".[15]

Ziel der GPL ist dabei nicht, dass jede Bearbeitung von Software den Bestimmungen der GPL unterworfen wird. Die Verbreitung der Open Source Software mit proprietären Lizenzbedingungen erlaubt die GPL. Allerdings wird dies nur unter der Voraussetzung erlaubt, dass die Open Source Software und proprietäre Software voneinander getrennt weitergegeben werden und somit kein abgeleitetes Werk entsteht. Die unter der GPL stehende Software kann damit nicht verkauft oder gegen Vergütung lizensiert werden.[16]

Die GPL berechtigt den Nutzer damit zur uneingeschränkten Weiterverbreitung der Software ohne Erhebung von Lizenzgebühren, Nutzung für jegliche Zwecke, Vervielfältigung, Bearbeitung und Weitergabe unveränderter und veränderter Versionen der Software. Dabei muss er jedoch einige Pflichten, wie z.B. das Beifügen des vollständigen Lizenztextes der GPL, Mitlieferung des Sourcecodes, sowie die Beachtung des Copylefts bei abgeleiteten Werken, erfüllen.[17]

Neben der GPL existieren noch weitere Lizenzen mit Copyleft-Effekt, z.B. die IBM Public License und die Common Public License.

2.3.2.2 Lizenzen mit beschränktem Copyleft-Effekt

Diese Lesser General Public License ist eine Variante, die zwischen Lizenzen mit strengem Copyleft-Effekt (GPL) und Lizenzen ohne Copyleft-Effekt (BSD) steht.

Die Weiterentwicklung der Software wird grundsätzlich den jeweiligen ursprünglichen Lizenzbedingungen unterstellt (Copyleft). Werden jedoch Modifikationen des Quellcodes in eigenen Dateien vorgenommen, können diese unter anderen, auch proprietären Lizenzbe-

[15] Vgl. BITKOM (2006), S.10
[16] Vgl. BITKOM (2006), S.10
[17] Vgl. BITKOM (2006), S.9f

dingungen weitergegeben werden. Es damit möglich Software unter mehreren Lizenzarten zu kombinieren, allerdings nicht uneingeschränkt.[18]

Die Lesser General Public License wurde speziell für Programmbibliotheken entwickelt, um freie Standardbibliotheken auch im kommerziellen Softwarebereich zu verbreiten. Dies wäre bei Anwendung der strengen Regelungen der GPL nicht möglich. Grundsätzlich gelten für die Weitergabe von unveränderten und veränderten LGPL-Bibliotheken dieselben Pflichten wie bei der GPL. Der Nutzer erhält u. a. das Recht, diese zu ändern und zu verbreiten. Bei Weitergabe veränderter Bibliotheken greift gleichfalls der strenge Copyleft-Effekt.[19]

Bezüglich des Verhältnisses zwischen zugreifendem Programm und LGPL-Bibliothek sind die Regelungen der Lizenz jedoch deutlich modifiziert. Programme, die auf eine LGPL-Bibliothek lediglich zugreifen, bleiben proprietär, wenn Software und Bibliothek unabhängig hiervon vertrieben werden. Da es sich hierbei um kein abgeleitetes Werk handelt und fällt es nicht unter das Copyleft.[20]

Anders stellt sich die Situation jedoch dar, wenn ein Programm mit einer LGPL-Bibliothek verlinkt (d.h. verknüpft) und gemeinsam vertrieben wird. Diese Art des Zusammenspiels zwischen Programm und LGPL-Bibliotheken führt zur Annahme eines neuen Datenwerks und damit zum Copyleft-Effekt, der die gesamte Softwareentwicklung umfasst.[21]

2.3.2.3 Lizenzen ohne Copyleft-Effekt

Die Lizenzbedingungen ohne Copyleft-Effekt enthalten weniger Pflichten als die zuvor dargestellten Lizenzbedingungen. Hinzu kommt, dass Veränderungen der Software vorgenommen und unter beliebigen, auch eigenen, Lizenzbedingungen verbreitet werden können. Eine Kombination mit proprietärer Software ist damit auch möglich.[22]

Der Nutzer erhält das Recht zur Vervielfältigung, Veränderung, zum Vertrieb von veränderten und unveränderten Versionen als Sourcecode oder im Objektcode. Die Vermarktung von Software unter eigenen, auch proprietären Lizenzen, ohne Offenlegung des Sourcecodes ist möglich. Im Gegenzug muss der Nutzer einige Pflichten erfüllen, wie beispielsweise die Weitergabe der Open Source Lizenz und Beibehalten des Copyright-Vermerks. [23]

Bekannte Beispiele für Lizenzen ohne Copyleft ist die BSD (Berkeley Software Distribution), Apache Software License und OpenLDAP Public License.

[18] Vgl. BITKOM (2006), S.11f
[19] Vgl. BITKOM (2006), S.12
[20] Vgl. BITKOM (2006), S.12
[21] Vgl. BITKOM (2006), S.12
[22] Vgl. BITKOM (2006), S.14
[23] Vgl. BITKOM (2006), S.14

2.3.3 Vorteile und Nachteile für Unternehmen

Bevor sich ein Unternehmen für die Nutzung von Open Source Software entscheidet, ist es wichtig sich mit den Chancen und Risiken von Open Source Software vertraut zu machen. Im Folgenden soll auf die wichtigsten Vor- und Nachteile von Open Source Software eingegangen werden. Abbildung 1 zeigt die am häufigsten genannten Beweggründe bei dem Wechsel von proprietärer Software zu Open Source Software.

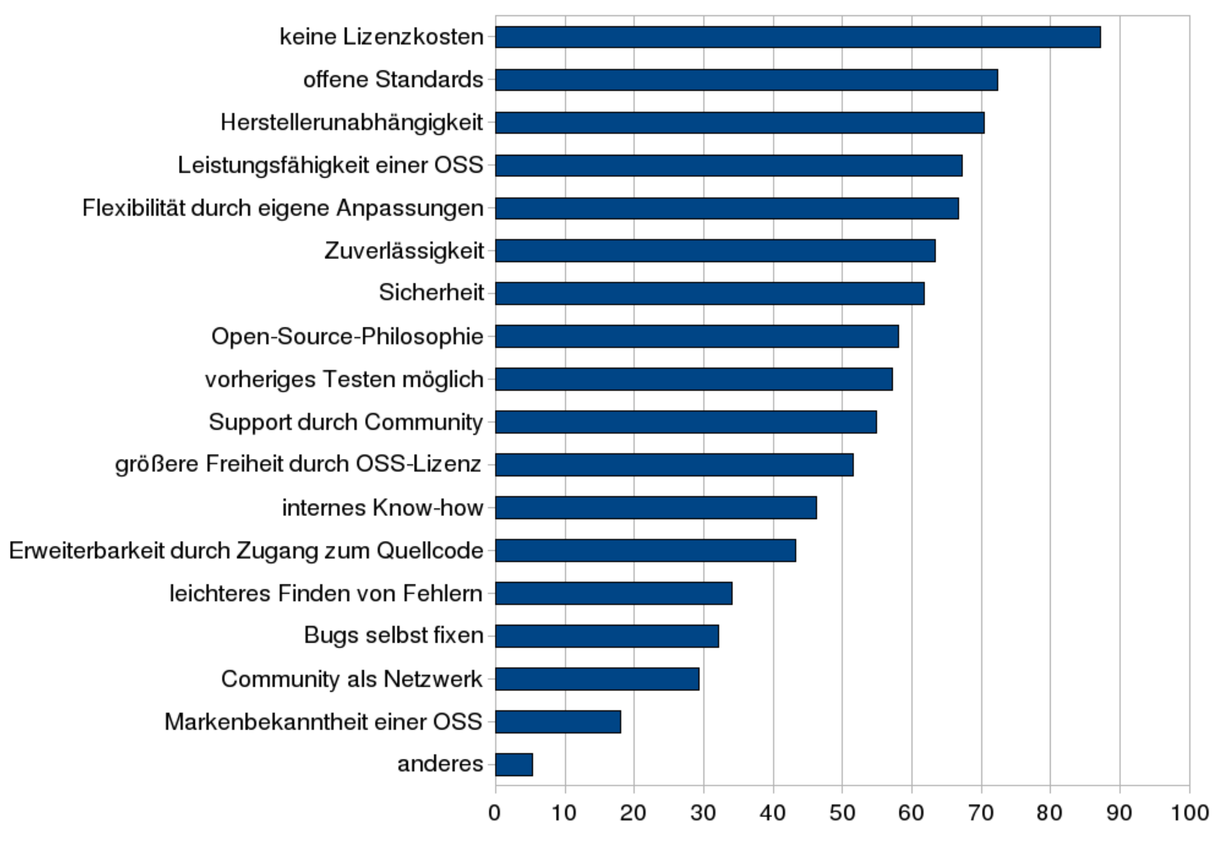


Abbildung 2: Beweggründe für den Einsatz von OSS[24]

Der mit Abstand am häufigsten genannte Grund ist der Wunsch Lizenzkosten einzusparen. Bei proprietärer Software sind die Kosten, die bei dem Unternehmen entstehen zu einem großen Teil durch diese Lizenzkosten bestimmt. Aus diesem Grund ist nachgewiesen, dass Open Source Produkte gegenüber proprietären Produkten erheblich geringere Kosten verursachen. Hierbei darf jedoch nicht vernachlässigt werden, dass Schulungskosten einen sehr wichtigen Teil der Kostenrechnung darstellen, besonders dann, wenn eine Migration angestrebt wird. Diese Problematik ist jedoch meist nur mit Endanwendern relevant, da z.B. Administratoren oftmals über umfangreiche Linux-Kenntnisse verfügen und somit keine Schulung bei der Einführung benötigen. Dies hängt mit der Offenheit der Systeme und der häufig sehr guten Dokumentation von Open Source Software zusammen.[25]

[24] Diedrich, O. (2009), S.4

[25] Heinrich, H. / Holl, F. / Menzel, K. / Mühlberg, J. / Schäfer, I / Schüngel, H. (2006), S.21

Als zweites wichtigstes Argument werden offene Standards genannt. Die von Open Source Produkten verwendeten Dateiformate und Datenaustauschstandards sind über die Verfügbarkeit des Sourcecodes per Definition offen gelegt. Damit ist eine höhere Kompatibilität und Interoperabilität von Open Source Produkten mit anderer Software garantiert, was zur Folge hat, dass entsprechende Schnittstellen mit geringerem Aufwand erstellt werden können.[26]

Des Weiteren zwingt Open Source Software im Gegensatz zu einigen kommerziellen Anbietern die Nutzer nicht in ein Abhängigkeitsverhältnis zu bestimmten Herstellern. Die Vielzahl an Nutzungsbedingungen, die oftmals bei kommerzieller Software vorhanden ist, muss als Anwender von Open Source Software nicht befürchtet werden. Im Gegensatz dazu wird bei Open Source viel Wert auf umfassende Freiheitsrechte im Bezug auf den Einsatz der Software gelegt.[27]

Neben den Vorteilen können jedoch auch einige Nachteile bei der Nutzung von Open Source Software auftreten. Oftmals können Nutzer von Open Source Software weder Gewährleistungs- noch Haftansprüche gegen die Entwickler der Software geltend machen. Die heutzutage häufig verwendeten Open Source Lizenzen übernehmen auch keine Garantie für die Funktionsfähigkeit der Software, was dazu führt, dass der Anwender das volle Risiko trägt.

Hinzu kommt, dass nur sehr selten von Seiten der Entwickler von Open Source Software Support für ihre Produkte angeboten wird. Teilweise müssen Anwender damit auf Dienstleistungen Dritter zurückgreifen. In den letzten Jahren passiert es jedoch immer öfter, dass Entwickler von Open Source Software das Angebot von passenden Dienstleistungen als Geschäftsmodell ansehen und Support kostenpflichtig anbieten.[28]

Es gibt noch eine Vielzahl an weiteren Vor- und Nachteilen von Open Source Software, die im Rahmen dieser Arbeit nicht weiter Untersucht werden können. Die folgende Tabelle stellt die wichtigsten Vor- und Nachteile zusammenfassend gegenüber.

[26] Renner, T. / Vetter, M. / Rex, S. / Kett, H. (2005) , S.17
[27] Renner, T. / Vetter, M. / Rex, S. / Kett, H. (2005) , S.16
[28] Renner, T. / Vetter, M. / Rex, S. / Kett, H. (2005) , S.17

Vorteile	Nachteile
Anpassbarkeit	Keine Gewährleistungsrechte
Wiederverwendbarkeit von Code	(Oft) kein Support durch Entwickler
Höhere Produktqualität	Höherer Schulungsaufwand
Anbieterunabhängigkeit	Ungewisse Weiterentwicklung
Höhere Sicherheit	Applikationen teilweise nicht erhältlich
Offene Standards	Teilweise mangelnde Interoperabilität mit kommerzieller Software
Keine Lizenzkosten	

Tabelle 2: Vor- und Nachteile von OSS[29]

2.4 PDF und technischer Hintergrund

2.4.1 Was ist das PDF Format?

Portable Document Format (PDF) ist ein Datenformat. Das Format kann ein Dokument unabhängig von Software, Hardware oder Betriebssystem darstellen. Dies bedeutet, dass das Dokument plattformunabhängig genau in der Form und Formatierung dargestellt wird, wie es zu Beginn der Formatierung aussah. Wenn also zum Beispiel ein MS Word Dokument in ein PDF Dokument konvertiert wird, wird das MS Word Dokument als PDF einfach originalgetreu dargestellt.[30]

2.4.2 Geschichte und Entstehung

Das PDF wurde in den 90er Jahren von der Firma Adobe Systems entwickelt. Die erste Version des Formats wurde im Jahr 1993 veröffentlicht und seitdem wird das Format stetig weiterentwickelt.[31]

Bei der Einführung des Datenformates wurden sowohl das PDF Erzeugungswerkzeug, als auch der PDF Reader kostenpflichtig angeboten. Zu Beginn war das PDF Format daher nicht erfolgreich. Erst als die Strategie geändert wurde, startete der Siegeszug von PDF. Von nun an konnte der PDF Reader kostenlos erworben werden und somit und bereit nach kurzer Zeit stieg die Anzahl der Nutzer.[32] Da sich das PDF Format im Laufe der Zeit als Standard

[29] Renner, T. / Vetter, M. / Rex, S. / Kett, H. (2005) , S.17
[30] Vgl. Bienz, T./Cohn, R./Meehan, J.R. (1996), S. 29
[31] Vgl. Prepressure (2013)
[32] Vgl. Prepressure (2013)

durchgesetzt hat, wurden immer mehr alternative Anwendungstool entwickelt. Heutzutage gibt es sowohl kostenpflichtige, als auch kostenlose Software die genutzt werden kann.

2.4.3 Funktionsumfang

Eine Begrenzung der Seitenzahlen eines PDF Dokumentes gibt es generell nicht. Dies bedeutet, dass eine PDF-Datei auch mal tausend Seiten umfassen kann. Inhaltlich kann ein solches Dokument viele unterschiedliche Formen wie zum Beispiel Text, Grafiken oder auch Hyperlinks enthalten.[33]

PDF ist eine vektorbasierte Sprache zur Seitenbeschreibung. Alle Informationen in einem PDF werden als Objekt angesehen und abgespeichert. Ein solches Objekt kann zum Beispiel die Schriftgröße und Form sein, aber auch eine Seitenanzahl auf einer Seite. Je nach Umfang des PDF Dokumentes kann eine Datei mehrere 1000 Objekte enthalten.[34] Ein Vorteil dieser Vektorbasierten Seitensprache und der originalgetreuen Darstellung des PDF Dokumentes ist zum Beispiel, dass selbst wenn eine spezielle Schriftart auf dem Rechner installiert ist und verwendet wird, diese Schriftart in der PDF Datei noch genauso aussieht wie bei der Erstellung. Wenn man im Vergleich dazu Word Dateien verschickt, greift eine solche Word Datei auf die auf dem Rechner installieren Schriftarten zurück. Wenn eine bestimmt Schrift nicht vorhanden ist, wird der Text dann einfach in einer der bekannten Schriften dargestellt und das komplette Word-Dokument wird durch den Schriftwechsel verzerrt.

Eine PDF Datei kann durch JavaScript Programmierung weitere Funktionen integrieren. So können zum Beispiel Lesezeichen oder Kommentare erstellt werden. So kann zum Beispiel auch ein Navigationsmenü in ein PDF Dokument integriert werden. Zusätzlich ist es möglich Formulare zu erstellen, die dann auch direkt online ausgefüllt und gespeichert werden können.[35] Der Vorteil hier ist, dass nun ehemals handschriftlich ausgefüllte Dokumente sofort elektronisch ausgefüllt und digital gespeichert sind. Dies ist ein erfolgreicher Schritt zu einem papierlosen Büro.[36]

Eigene Sicherheitsmechanismen können verwendet werden um ein Produkt zu schützen. So kann zum Beispiel eingestellt werden, ob Text später im PDF Dokument als Text ausgegeben wird und somit kopiert werden kann oder nicht. Wenn der Text nicht erfasst wird, so handelt es sich bei dem Dokument also wie um ein Bild. Es ist jedoch grundsätzlich möglich, bestimmte Objekte aus einer PDF Datei herauszukopieren und diese in eine andere Anwendungssoftware wie zum Beispiel MS Word zu integrieren. Hier können dann diese Dokumente beliebig weiterverarbeitet werden. Dies gibt natürlich nur, wenn der Ersteller des PDF Dokumentes dies zugelassen hat. Es ist auch möglich, die Kopierfähigkeit und die Texterken-

[33] Vgl. Bienz, T./Cohn, R./Meehan, J.R. (1996), S. 29
[34] Vgl. Schubert, T. (2002), S. 8
[35] Vgl. Adobe (2007), S. 35 ff
[36] Vgl. Eggeling, T. (2008)

nung zu deaktivieren. Durch die interne Suchfunktion in PDF-Reader Software ist es möglich, schnell auf bestimmte Informationen zuzugreifen.

Ein PDF Dokument kann verschlüsselt werden. Für diesen Schutz kann entweder eine 40-Bit oder eine 128 Bit Verschlüsselung gewählt werden. Ebenso ist es möglich, eine Datei mit einem Passwortschutz einzurichten. Hier kann dann auf das Dokument nur mit dem richtigen Passwort zugegriffen werden.[37]

Das PDF-Format wurde im Jahr 2008 unter dem ISO-Standard 32000 veröffentlich.[38]

Das PDF-Format bietet auch noch spezielle Sonderformate an. PDF/A ist ein Beispiel zur elektronischen Archivierung von Dateien. Es gibt teilweise die Anforderung, dass Dokumente über einen gewissen Zeitraum hinweg aufgehoben werden müssen. Um die Unveränderlichkeit der Daten zu gewährleisten, muss dazu der Standard von PDF/A genutzt werden.[39] Zusätzlich gibt es auch noch ein PDF/UA Standard. Dieser behandelt den barrierefreien Aufbau eines PDF-Dokumentes. Relevant wieder dieser Standard also für zum Beispiel Formular, die von öffentlichen Ämtern genutzt werden. Außerdem ist noch das Format PDF/E ein relevantes Format, dass für technische Dokumente und Abhandlungen genutzt werden kann.[40]

[37] Vgl. Schubert, T. (2002), S. 25
[38] Vgl. Adobe (2008), S. 5
[39] Vgl. Heiermann, C. (2013)
[40] Vgl. Drümmer, O. (2011), S. 11ff

3 Praktische Abhandlung

3.1 Kriterienkatalog in der Praxis

Der Kriterienkatalog dieses Projekts wurde zusammen mit einem Fachmann eines großen Versicherungsunternehmens, sowie entlang der Software „pdfFactory“ erstellt. „pdfFactory“ bietet sich aufgrund diverser Gründe als Vergleichsbasis gegenüber den Open Source Produkten an. Erstens, „pdfFactory“ wird in großen Unternehmen eingesetzt, zweitens bietet pdfFactory nahezu alle für ein Unternehmen benötigten Anforderungen an ein PDF Erzeugungstool an. Dementsprechend ist die Grundlage für die Kriterien des Kriterienkatalogs die Einschätzung des Fachmanns, sowie „pdfFactory“. Die Kriterien wurden in nicht-funktionelle und funktionelle Anforderungen unterteilt und daraus wurde folgender Kriterienkatalog inklusive Gewichtung erstellt:

Unfunktionale Anforderungen

Kriterium	Gewichtung
Hersteller	0
Website	0
Lizenz	5
Aktive Community (Letztes Update)	5
Reifegrad	5
Systemanforderung (mindestens Windows 7)	5
Sprache (mindestens Deutsch und Englisch)	5
Umfang (Konverter, Editor, etc.)	3
Technische Basis (Ghostscript, etc.)	3
PDF-Spezifikation (unterstützte PDF-Formate)	1
Kosten für Produkt	5
Folgekosten für Support	5

Tabelle 3: Unfunktionale Anforderungskriterien

Funktionale Anforderungen

Kriterium	Gewichtung
PDF/A	5
PDF/UA	3
PDF/X	3
Kompatibilität mit anderen Programmen (Add-ins)	1
Direkt-Konvertierung Microsoft Office 2010	2
Direkt-Konvertierung OpenOffice	2
Personalisierung (Serienbrief, etc.)	2
PDFs zusammenfügen (merge)	5
PDFs aufteilen (split)	5
Briefpapier / Wasserzeichen	3
Schwärzen von Informationen	5
Verschlüsselung	5
Zugriffsschutz	4
Digitale Signatur	3
Automatisierte Vorgänge	1
OCR	1
Offline nutzbar	5
Intuitive Bedienung	4
Direkter E-Mail-Versand	2
Vorschau	1
Benutzerhilfe	3
Dokumentation	3
Druckvorlagen	2
Maximale Dokumentenauflösung (in dpi)	2
Komprimierung der Grafikauflösung	5

Tabelle 4: Funktionale Anforderungskriterien

Als wesentliche unfunktionale Kriterien und damit als KO-Kriterien wurden eine aktive Community, der Reifegrad der Software, die Systemanforderung und die Sprache definiert. Eine aktive Community und der Reifegrad der Software sind speziell bei Open Source Produkten sehr wichtig. Das Engagement für die Weiterentwicklung kommt aus der Entwicklungscommunity. Nur wenn diese aktiv am Open Source Produkt arbeitet, kann davon ausgegangen werden, dass die Software in Zukunft weiter funktionsfähig bleibt und verbessert und weiterentwickelt wird. Eine wesentliche funktionalen Anforderungen ist zudem die Unterstützung

des Formats PDF/A. Dieses Format wird zur elektronischen Archivierung verwendet. Sowohl PDF/A, als auch die Funktion Split and Merge sind KO-Kriterien. Das Zusammenfügen und Teilen von PDF-Dateien wird verwendet, um zum Beispiel einen gewissen Auszug aus einem Dokument zu extrahieren oder mehrere PDFs zu einem PDF zusammenzufügen und übersichtlicher für den Bearbeiter oder Empfänger des Dokuments zu machen. Ein weiteres KO-Kriterium ist die Anforderung Inhalte in einem PDF-Dokument schwärzen zu können. Dies wird genutzt, wenn zum Beispiel sensible Daten nicht an Dritte gelangen dürfen oder ein Dokument anonymisiert werden sollen. Ebenso sind die Möglichkeit Grafiken zu komprimieren und die Offline-Nutzung weitere relevante KO-Kriterien. Die Komprimierung ist wichtig, um den Speicherbedarf möglichst gering und damit kostengünstig zu halten. Die Offline-Nutzung hingegen soll gewährleisten, dass das Produkt rund um die Uhr genutzt werden kann.

Neben den Ausschluss- oder auch KO-Kriterien sind auch zusätzliche Kriterien vorhanden. So gibt es unfunktionale Zusatzkriterien, die keinen Einfluss auf die Bewertung eines Produktes haben und nur zur Information dienen, sind „Hersteller“ und „Website“.

Weitere Kriterien mit mittlerer Gewichtung sind der Umfang der Software, die technische Basis und die unterstützen PDF-Formate. Dabei gilt, je mehr PDF-Formate unterstützt werden, desto geringer der Aufwand der technischen Basis und je größer der Umfang der Software ist, desto besser die Bewertung. Diese Kriterien geben im groben eine erste Übersicht darüber, ob ein oder mehrere Tools nötig sind um den kompletten Anforderungskatalog zu erfüllen.

Zusätzliche funktionale Kriterien sind die Unterstützung der Formate PDF/UA und PDF/X. PDF/UA ist ein Format, dass PDF-Dokumente barrierefrei für zum Beispiel behinderte Menschen zur Verfügung stellt. Das Format PDF/X kann genutzt werden, um Druckvorlagen zu erstellen.

Die Kompatibilität zu anderen Programmen, die Direkt-Konvertierung von Microsoft Office und OpenOffice Dokumenten sind Kriterien, die keine hohe Gewichtung einnehmen. Mittlerweile bietet sowohl Open Office als auch Microsoft Office die Funktion der Direktkonvertierung aus dem Programm heraus an. Es ist in machen PDF-Programmen zudem möglich, die Erstellung von PDF-Dateien zu personalisieren. Dies könnte zum Beispiel dazu genutzt werden Serienbriefe automatisch zu erstellen. Diese Anforderung ist jedoch auch nur mit einer geringen Gewichtung im Vergleich berücksichtigt.

Weitere Zusatzkriterien mit mittlerer bis hoher Gewichtung sind die Möglichkeit Wasserzeichen in PDF Dokumenten zu setzen, eine digitale Signatur und einen Zugriffsschutz einzurichten. Die digitale Signatur soll die Echtheit eines Dokumentes bestätigen und der Zugriffsschutz soll unerlaubtem Zugriff auf das PDF-Dokument verhindern.

Features mit geringer Gewichtung sind unter anderem der direkte E-Mail-Versand von Dokumenten. Hiermit könnten Arbeitsschritte optimiert und Medienbrüche vermieden werden. Die automatische Texterkennung, auch unter OCR bekannt, ist ebenso eher als unwichtig markiert. Weitere wichtige Anforderungen an eine Software sind eine Benutzerhilfe und eine Dokumentation vom von der Software. Beides erleichtert dem Nutzer den Umgang mit dem Tool. Zum Ende der Kriterienliste ist noch die Möglichkeit aufgeführt, die Druckauflösung (dpi) einstellen zu können.

Anhand dieses Kriterienkatalogs wird nun Open Source Software, sowie pdfFactory analysiert und miteinander verglichen

3.2 Analyse des aktuell genutzten Tools pdfFactory

Das unter proprietärer Lizenz stehende pdfFactory steht als Gegenpart zu den Open-Source Lösungen. Anhand des in 3.1 erstellten Kriterienkatalogs wurde die Software getestet und bewertet. Das Ergebnis des Tests ist somit vergleichbar mit den Open Source Lösungen.

Das Programm „pdfFactory“ von der Context-GmbH wird unter anderem bei einer großen Versicherung in Deutschland eingesetzt. Das Tool erfüllt alle funktionalen Anforderungen mit einer guten Bewertung. Jedoch ist „pdfFactory“ mit 100€ pro Lizenz ein sehr teures Produkt. Der Preis ist abhängig davon, ob eine Server Edition gekauft wird und wie viele User auf das Produkt zugreifen können soll. Eine vollfunktionsfähige Testversion ist kostenfrei für 30 Tage erhältlich. Das Tool „pdfFactory“ wird als Standardversion und als „pdfFactory pro“, mit mehr Funktionen, angeboten. Wie aus dem Kriterienkatalog zu erkennen ist, bietet pdfFactory viele Funktionen, die die Arbeit im Büroalltag unterstützen. Neben dem Zusammenfügen und der Trennung einzelner PDFs, verfügt das Tool über weitere wichtige Funktionen, die zum Beispiel bei der Arbeit mit sensiblen Daten notwendig sind. So ist es möglich bestimmte Informationen zu schwärzen, erstellte PDFs zu verschlüsseln und mit Zugriffsrechten auszustatten. Ebenso ist es möglich Bilder und Grafiken zu skalieren und direkt per E-Mail zu versenden, PDF-Dateien digital zu signieren und Wasserzeichen zu setzen. Außerdem ist pdfFactory mit Microsoft Office und anderen Office Programmen kompatibel, da pdfFactory als Drucker eingerichtet wird. Außerdem ist dieses Tool mit PrintFine kompatibel. Mit PDF/A 1b, 2b und 3b werden die wichtigsten Archivformate unterstützt.

Die Auswertung zum Tool kann im Anhang „Anhang 1 – Auswertung pdf Factory“ gefunden werden.

Fazit: Zusammenfassend ist zu sagen, dass „pdfFactory“ alle notwendigen Funktionen, die im Umgang mit sensiblen Daten benötigt werden, bietet. Außerdem werden noch zusätzlich Funktionen angeboten, die den Arbeitsalltag unterstützen. Damit eignet sich das Produkt

überaus für die Verwendung im Büroalltag, insbesondere für Versicherungsunternehmen und anderen Unternehmen, die alltäglich mit sensiblen Daten arbeiten. Jedoch ist das Produkt, wie vorher erwähnt, sehr teuer.

3.3 Analyse und Beschreibung der OS Tool Tests

3.3.1 Beschreibung des Vorgehens

Anhand der Kriterien, die in den vorherigen Kapiteln beschrieben wurde, wurden mehrere Produkte getestet. Hierbei wurde die Software jeweils nach den Kriterien heraus betrachtet und bewertet.

Im ersten Schritt dieser praktischen Analyse wurden Produkte ausgewählt, die für einen solchen Vergleichstest in Frage kommen würden. Diese Produkte wurden durch Internetrecherchen gefunden und in eine gemeinsame Liste aufgenommen. Nach Abschluss der Recherche, wurde diese Vorauswahl genauer betrachtet und einige Produkte wieder aussortiert. Ein Kriterium für das Aussortieren war zum Beispiel die Verfügbarkeit des Produktes. Einige PDF-Tools können nur Online genutzt werden. Da das Auftragsunternehmen mit sensiblen Personenbezogenen Daten agiert, darf eine solche unsichere Schnittstelle nicht genutzt werden. Ein weiteres Kriterium für die Aussortierung war auch die nutzbare Plattform des Produktes. Wenn eine Software zum Beispiel nicht auf einer Windows Plattform genutzt werden kann, wurde das Produkt ebenso sofort aussortiert. Da das Versicherungsunternehmen hauptsächlich Windows 7 PCs im Betrieb hat, ist die Plattform ein entscheidendes Kriterium. Die Erstauswahl an Produkten ist im Anhang 4 nachzulesen.

Nach der ersten Vorauswahl und der Aussortierung einiger Produkte blieben zehn Produkte übrig. In dieser Auswahl wurden sowohl GPL-Lizenz-Produkte integriert, als auch proprietäre Freeware Produkte.

Software	Lizenzmodell
Altarsoft PDF Converter	Freeware
Broadgun pdfMachine	Freeware
Foxit Reader	Freeware
FreePDF	Freeware
Inkscape	GPL
jpdftweak	AGPL
PDF Split and Merge	GPL
PDF24 Creator	Freeware
PDFCreator	GPL
pdftk	GPL

Tabelle 5: Auswahl möglicher PDF-Produkte

3.3.2 Testergebnisse

Für die Durchführung der Tests wurde die jeweilige Software auf einen Computer heruntergeladen. Dadurch konnte die Software auch tatsächlich im Betrieb getestet werden. Informationen zu den Testplattformen, Softwareversionen und Weiterem können im Anhang 2 eingesehen werden.

3.3.2.1 Altarsoft PDF

Altarsoft ist ein Anbieter von verschiedenen Programmen, die alle unter einer kostenlosen proprietären Lizenz laufen. Von diesem Anbieter gibt es drei Produkte, die im Funktionsumfang etwas mit PDF zu tun haben. Zum einen gibt es einen Reader, der PDF-Dokumenten anzeigen kann. Es gibt einen sogenannten PDF Converter, der verschiedene Formate konvertiert. Hier gibt es sechs unterschiedliche Konvertierungsmöglichkeiten (pdf to images, pdf to text, text to pdf, images to pdf und split pdf). Die dritte Software nennt sich PDF Split Files und ermöglich das Trennen von einem PDF-Dokument in mehrere Dokumente.

Das Programm bietet mehrere Sprachen für die Nutzung an. Jedoch sind die Programme noch nicht in Deutsch erhältlich. Eine englische Version hingegen ist jedoch erhältlich.

Auf der Webseite des Herstellers beziehungsweise des Entwicklers sind nur sehr wenig Informationen zu den eigenen Produkten aufgeführt. Es gibt eine Kontakt-E-Mailadresse, bei der man sich bei Bedarf melden kann. So wird zum Beispiel auch kein Benutzerhandbuch

zur Verfügung gestellt. Dieses wird jedoch nicht wirklich benötigt, da die Benutzung durch den geringen Funktionsumfang sehr intuitiv ist und notfalls auch einfach durch Ausprobieren genutzt werden kann.

Fazit: Insgesamt bieten die drei Programme nur die Basisfunktionen für die Nutzung von PDF-Dokumenten an. Als eigenständige Software ist der Funktionsumfang des Produktes nicht ausreichen. Da die einzelnen Funktionen jedoch immer nur in einzelnen Programmen abgebildet sind, würde sich diese Software als Zusatzkomponente, zum Beispiel als einzelnes Split-and-Merge-Programm anbieten.

3.3.2.2 Broadgun pdf Machine

Beim Produkt Broadgun pdf Machine handelt es sich um eine proprietäre Lizenz. Es gibt verschiedene Versionen der Software mit unterschiedlichem Funktionsumfang. Insgesamt werden bereits bei der einfachsten Softwarevariante viele der Anforderungen abgedeckt. Die Anwendung ist in zwei Funktionen, grob zusammengefasst der Erstellung und der Bearbeitung, unterteilt.

Das Programm ist mehrsprachig nutzbar. Broadgun pdf Machine bietet eine Kompatibilität mit anderen Softwareprodukten an. So kann in einer erweiterten Version eine Anbindung an zum Beispiel ZUGFeRD erstellt werden. ZUGFeRD ist eine automatisierbare Rechnungstellungssoftware. Einfachere Anforderungen wie das Teilen und Zusammenfügen oder auch Split and Merge genannt wird ebenso erfüllt wie das Angebot Verschlüsselung und einen Zugriffsschutz einzurichten. Ein besonderes Feature ist auch die automatische Texterkennung innerhalb der PDF-Dokumente. Bei der Speicherung des Dokumentes kann eine Komprimierung des Dokumentes durchgeführt werden. Ebenso ist es möglich, je nach Drucker die Druckauflösung einzustellen.

Die Software und die Benutzermaske kann individuell an den Benutzer und die jeweiligen Anforderungen angepasst werden. Vom Hersteller wird eine Benutzerhilfe zur Unterstützung angeboten.

Die Funktion „Schwärzen von Informationen" ist im Anforderungskatalog erwünscht. Dies wird jedoch nicht von der Software angeboten.

Der breite Katalog der angebotenen Funktionen von Broadgun pdf Machine ist kostenpflichtig. Je nach Version und Anzahl der Benutzer steigen die Kosten für die Lizenzen. Eine Lizenz für die einfachste Version kostet circa 70 Euro. Eine preisliche Staffelung für mehr genutzte Lizenzen ist vorhanden.

Fazit: Die Software Broadgun PDF Machine hat ein umfangreiches Funktionsangebot das die Anforderungen weitgehend abdeckt. Jedoch handelt es sich bei dieser Software nicht wie

zu Beginn gedacht um eine proprietäre Freeware, sondern um eine kostenpflichtige Software. Es lässt sich daher die Frage stellen, ob die Software überhaupt noch innerhalb dieses Vergleiches berücksichtigt werden kann.

3.3.2.3 Foxit Reader

Die Funktionen vom Foxit Reader unterstützen einen großen Bereich des Anforderungskataloges. Die Lizenz für das Produkt ist eine proprietäre Lizenz. Als Privatnutzer kann man die Software kostenlos erwerben und nutzen. Für Unternehmen wird die Nutzung jedoch kostenpflichtig.

Das Design der Software ist ähnlich zu den Microsoft Office Produkten. In der Menübar sind alle wichtigen Funktionen aufgeführt und der Nutzer findet sich schnell und einfach zurecht. Foxit Reader läuft sowohl auf Windows und iOS. Zusätzlich gibt es eine Android Version, worüber man auf mobilen Endgeräten PDF Dokumente anschauen und bearbeiten kann.

Es gibt verschiedene Plug-Ins, die in den Funktionskatalog von Foxit-Reader integriert werden können. So können zum Beispiel weitere Anforderungen über diese Alternative doch noch erfüllt werden. Es gibt keine Funktion für den Zugriffsschutz (z.B. Passwortgeschützt) direkt auf das Dokument. Jedoch ist möglich, ein PDF-Dokument zu signieren und zu verifizieren. Es können Bilder mit unterschiedlicher Transparenz in ein Dokument eingefügt und gespeichert werden. Durch diese Funktion, kann man ein Wasserzeichen zum Beispiel der Firma integrieren. Aus dem Programm heraus können direkt E-Mails mit dem jeweiligen Dokument als Anhang versendet werden.

Eine Benutzerhilfe wird vom Hersteller zur Verfügung gestellt. Diese ist auch sinnvoll, da das Programm einen umfangreichen Funktionskatalog anbietet.

Eine automatische Texterkennung wird nur in einer erweiterten und damit kostenpflichtigen Version angeboten. Es können keine Vorgänge automatisiert werden. Die Funktionen Split and Merge werden vom Foxit Reader nicht unterstützt.

In Foxit Reader ist möglich, Druckvorlagen auszufüllen. Diese ausgefüllten Dokumente können dann auch mit den Einträgen gespeichert werden. So wäre es zum Beispiel möglich, ein Formular bereits mit bestimmten allgemeinen Daten, die bei jedem Formular gleich bleiben, vorauszufüllen.

Fazit: Die Funktionen die das Programm anbietet sind umfangreich. Die Basisanforderungen sind erfüllt und im normalen Gebrauch von PDF-Dokumenten bleiben wenig Wünsche offen. Da es sich bei diesem Programm aber um eine proprietäre Lizenz handelt, ist der kommerzielle Gebrauch für Unternehmen kostenpflichtig.

3.3.2.4 FreePDF

FreePDF wird unter einer proprietären Lizenz kostenfrei angeboten. Der Entwickler erwähnt auf der Produktwebseite mehrmals, dass jede Nutzung, ob privat oder kommerziell, kostenlos ist. Somit können auch Unternehmen die Software kostenfrei einsetzten.

Es gibt ein Forum für Fragen zur Nutzung der Software. Dieses wird auch recht regelmäßig genutzt. Trotz der aktiven Nutzung, scheint es sich bei der Softwareentwicklung eher um ein kleineres Projekt zu handeln. Es wird kein kostenloser Support angeboten.

FreePDF ist nur auf Windows nutzbar. Es werden mehrere Sprachen für die Nutzung angeboten. Die Programmierbasis ist Ghostscript.

Der Funktionsumfang der Software ist im Vergleich zu anderen Produkten mittelmäßig aufgestellt. Die Bedienung erscheint im ersten Moment etwas kompliziert, jedoch wird das Konzept des Programmaufbaus nach kurzer Zeit schlüssig.

Die Software kann über die Druckfunktion als Druckeralternative direkt angesprochen werden. So können dann hiermit direkt PDF-Dokumente aus verschiedenen Programmen heraus erstellt werden. Dokumente können über die Merge-Funktion zusammengefügt werden. Zusätzlich gibt es eine Funktion der Verschlüsselung von Dokumenten. Einen Zugriffsschutz oder einen Passwortschutz gibt es jedoch nicht. Es wird eine Benutzerhilfe vom Entwickler online zur Verfügung gestellt.

Eine Split-Funktion gibt es in FreePDF nicht. Eine Personalisierung des Systems auf den Nutzer ist nicht möglich. Die Anforderung der Informationsschwärzung in PDF-Dokumenten wird ebenso nicht unterstützt. Eine automatische Texterkennung gibt es nicht. Kompromierung von Daten bzw. die Dokumentenauflösung können ebenso nicht eingestellt werden. Zwar ist die Verschlüsselung eines Dokumentes möglich, jedoch kann ein Dokument nicht digital signiert oder verifiziert werden. Der Entwickler hat scheinbar eine Kooperation mit einem weiteren Softwareentwickler. Die Anforderung für das Einfügen eines Wasserzeichens kann daher eventuell über ein Zusatzprogramm erfüllt werden.

Fazit: Das positive an der Software FreePDF ist, dass sie trotz proprietärer Lizenz auch im Unternehmensumfeld komplett kostenlos ist. Es ist jedoch möglich, den Entwickler mit Spenden zu unterstützen. Der Funktionsumfang der Software ist nicht allumfassend, jedoch werden mehrere wichtige Anforderungen erfüllt. Dieses Produkt würde in Kombination mit einem weiteren Ergänzungsprodukt genutzt werden.

3.3.2.5 Inkscape

Inkscape ist eine Open-Source Software zur Bearbeitung und Erstellung zweidimensionaler Vektorgrafiken, welche unter der GPL lizensiert ist. Seit der Version 0.46 wird jedoch das Importieren und Exportieren von PDF-Dokumenten unterstützt.

Die Inkscape Website ist sehr gut aufgebaut und bietet eine große Anzahl an nützlichen Informationen und Hilfestellungen. Es gibt diverse Kommunikationsplattformen wie z.B. einen IRC-Chat, Mailinglisten und ein Forum, die genutzt werden können um Fragen zu stellen und in Kontakt mit der Community zu kommen. Des Weiteren gibt es Tutorials, Video-Tutorials, ein FAQ und diverse Bücher/Handbücher, die kostenlos zur Verfügung gestellt werden.

Inkscape unterstützt Windows, Mac und Linux Systeme. Es ist sowohl auf Deutsch, als auch auf Englisch und vielen weiteren Sprachen verfügbar.

Da Inkscape eine Software zur Erstellung und Bearbeitung von Vektorgrafiken ist, sind sehr viele Funktionen verfügbar, die nicht bei der Arbeit mit PDF-Dokumenten genutzt werden. Trotz dessen bietet Inkscape einige Funktionen wie z.B. das Erstellen eines Wasserzeichens, das Schwärzen von Informationen, digitale Signaturen und das Komprimieren der Grafikauflösung. Es fehlen jedoch wichtige Funktionen, die bei der Arbeit mit PDF-Dokumenten benötigt werden.

Fazit: Inkscape ist eine sehr facettenreiche Software, die viele Funktionen bietet. Bei der Arbeit mit PDF-Dokumenten fehlen jedoch einige wichtige Funktionen, sodass es im Unternehmenskontext eher ungeeignet ist.

3.3.2.6 jpdf Tweak

jpdf Tweak ist eine Java Swing Applikation, die unter der AGPL lizensiert ist. Die Software kann somit privat als auch in Unternehmen genutzt werden. Die Lizenz ermöglicht das Einsehen und Verändern des Quellcodes, was ein Vorteil für Unternehmen sein kann, die eigene Veränderungen an der Software vornehmen möchten.

Die Website von jpdf Tweak ist sehr einfach gehalten und bietet nur eine kurze Übersicht über das Produkt und die Download-Links für den Installer bzw. den Quellcode. Außerdem wird angegeben, dass bei Anregungen und Bug-Reports eine E-Mail an den Entwickler gesendet werden kann.

jpdf Tweak kann sowohl auf Windows als auch auf Linux Rechnern installiert werden. Hierbei sind keine speziellen Systemanforderungen auf der Website zu finden. Die Software ist nur in englischer Sprache verfügbar.

Nach der Installation stehen dem Benutzer einige interessante und nützliche Funktionen zu Verfügung. Die Software ermöglicht das Zusammenfügen mehrerer PDF-Dokumente, das Aufteilen eines Dokumentes, das Hinzufügen eines Wasserzeichens und das Rotieren einzelner Seiten. Außerdem können Lesezeichen und Seitenübergänge sowie Anhänge hinzugefügt, Dokumente verschlüsselt und signiert und die Metadaten editiert werden.

Fazit: jpdf Tweak bietet für eine kostenlose Open-Source Software viele gute Funktionen für die Arbeit mit PDF-Dokumenten. Da die AGPL Lizenz verwendet wird, können eigene Änderungen am Quellcode vorgenommen werden, um beispielsweise neue Funktionen hinzufügen, oder um die Software zu personalisieren. Es fehlen jedoch einige Funktionen, die andere Programme in den Tests bieten konnten.

3.3.2.7 PDF Split and Merge (SAM)

PDF Split and Merge wird unter der Open Source Lizenz GPLv2 vertrieben. Die Entwicklungscommunity der Software ist aktiv und es werden regelmäßig Updates entwickelt. Momentan ist die Version 2.2.4 verfügbar. Innerhalb der kommenden Monate soll die Version 3.0 herausgebracht werden. Die Systemanforderungen für Windows sind gering und die Software ist mehrsprachig erhältlich. Die Programmierung des Systems wird mit Java durchgeführt.

Es kann eine kostenlose Version genutzt werden. Zusätzlich wird noch eine Enhanced Version mit zusätzlichen Funktionen angeboten. Diese Lizenz der Enhanced Version kann für 7,99 Euro erworben werden. Für Support gibt es ein Forum und ein Wiki. Zusätzlich kann an das Projekt eine E-Mail mit einer Supportanfrage geschickt werden.

Wie der Name des Programmes bereits verspricht, können PDF-Dokumente geteilt beziehungsweise getrennt werden. Hierfür können sich auch wiederholende Vorgänge automatisiert werden.

Da es sich bei PDF Split and Merge um ein Nischenprodukt handelt, ist der Funktionsumfang relativ klein. Viele der Anforderungen werden daher mit diesem Produkt nicht erfüllt.

Die Bedienung des Systems ist durch den geringen Funktionsumfang einfach. Es gibt eine Benutzerhilfe, die online zur Verfügung gestellt wird. Zusätzlich dazu kann die Dokumentation des Open Source Projektes ebenso online abgefragt werden.

Fazit: PDF Split and Merge bietet nur einen sehr geringen Funktionsumfang. Der Reifegrad und die Community lassen auf weitere Upgrades in der Zukunft hoffen. Durch die spezielle Anwendung des Produktes könnte es als Funktionsergänzung genutzt werden.

3.3.2.8 PDF24Creator

PDF24 Creator ist eine Freeware, die von der Firma geek Software angeboten wird. Wie auf der Website zu lesen ist, kann die Software sowohl von privaten Benutzern als auch Unternehmen komplett kostenfrei genutzt werden. Ein möglicher Nachteil dieser Freeware ist, dass der Quellcode nicht bearbeitet werden darf. Somit sind unternehmensinterne Änderungen und Anpassungen nicht erlaubt und strafbar.

Die Website erweckt einen professionellen und gepflegten Eindruck, was unteranderem auf der Integration eines Forums, FAQs, News-Feeds und eines ausführlichen Changelogs zurückzuführen ist. Der Changelog bietet eine Versionshistorie, welche Informationen zu allen Updates, Bugfixes und weiteren Verbesserungen liefert. Somit lässt sich sehr schnell und einfach verfolgen, welche Veränderungen an der Software vorgenommen worden sind und welche neuen Features verfügbar sind.

PDF24 Creator ist für Windows ab XP geeignet. Um das Ausrollen der Software in Unternehmen zu erleichtern, wird neben der regulären EXE-Installationsdatei eine MSI-Datei angeboten. Dies kann ein erheblicher Vorteil für Unternehmen darstellen, die eine Software auf eine sehr große Anzahl an Rechnern im Netzwerk installieren wollen. PDF24 Creator ist sowohl auf Deutsch, als auch auf Englisch verfügbar.

Die Funktionen von PDF24 Creator sind sehr umfangreich und erfüllen eine Vielzahl der Kriterien, die Unternehmen an ein PDF-Erzeugungstool haben. Die Software installiert einen virtuellen Drucker, der ein direktes Konvertieren aus Programmen wie Microsoft Office und anderen Programmen ermöglicht. Der Benutzer hat hierbei die Option, das Dokument in Standard-Formaten wie PDF/A oder PDF/X zu speichern. Es können Wasserzeichen hinzugefügt, Dokumente verschlüsselt und per Passwort geschützt werden, sowie digitale Signaturen hinzugefügt werden. Außerdem bietet PDF24 Creator eine grafische Oberfläche, die eine Vorschau des PDF-Dokuments anzeigt und ein Splitten und Mergen per Drag&Drop ermöglicht. Verschiedene Dokumente können so sehr einfach verbunden werden bzw. unnötige Seiten entfernt werden.

Fazit: PDF24 Creator ist ein Programm, mit sehr vielen nützlichen Funktionen für private Benutzer und Unternehmen. Es umfasst einen Großteil der wichtigen Anforderungen und ist im Unternehmenskontext komplett kostenlos. Da es sich jedoch nicht um eine Open-Source Anwendung handelt, ist ein Bearbeiten des Quellcodes nicht gestattet. Dies schränkt den Benutzer ein und kann somit für einige Unternehmen ein Ausschlusskriterium darstellen. In Anbetracht der Funktionalität und der kostenlosen Support-Möglichkeiten ist PDF24 Creator jedoch einer der klaren Favoriten in dieser Analyse.

3.3.2.9 PDFCreator

PDFCreator ist eine Open-Source Software von der Firma „pdfforge“, welche unter der AGPL lizensiert und angeboten wird. Aus diesem Grund wird auf der Website explizit angegeben, dass dieses Programm kostenlos privat zu Hause und in Unternehmen genutzt werden darf. Wie in der AGPL beschrieben, hat der Anwender Zugriff auf den Quellcode und kann diesen kompilieren, solange das Ergebnis unter der AGPL steht. Dies eignet sich besonders für Unternehmen, die selbstständig Änderungen im Quellcode vornehmen möchten, um beispielsweise neue Funktionen hinzuzufügen oder die Software zu personalisieren.

Die Website von PDFCreator wirkt sehr professionell und bietet verschiedene Support-Optionen wie ein Forum, ein sehr ausführliches Handbuch und FAQ. Außerdem wird ein Blog bereitgestellt, welcher Updates im Bezug auf neue Releases und weitere Informationen liefert.

PDFCreator ist nur unter Windows nutzbar und unterstützt alle Versionen ab Windows XP. Die Erzeugung der PDF-Dateien basiert auf der Ghostscript-API. Die Software kann auf Deutsch, Englisch und weiteren Sprachen genutzt werden.

Der Funktionsumfang von PDFCreator umfasst eine Vielzahl an nützlichen Features, die besonders in einem Unternehmenskontext von Vorteil sind. Da die Software wie ein virtueller Drucker im System fungiert, können Dateien direkt aus anderen Programmen wie Microsoft Office und Browsern wie Firefox und Internet Explorer konvertiert und erstellt werden. Hierbei können die Dokumente als PDF/A zur Langzeitarchivierung oder PDF/X gespeichert werden. Des Weiteren kann eine Verschlüsselung der Dokumente vorgenommen werden, wobei der Verschlüsselungsgrad vom Benutzer bestimmbar ist und ein Zugriffsschutz durch Vergabe eines Passworts erzeugt werden kann. Das Hinzufügen einer digitalen Signatur ist ebenfalls möglich.

Das Splitten und Mergen von Dokumenten ist mit PDFCreator nicht besonders einfach gelöst. Um mehrere Dokumente in ein Dokument zusammenzufügen, müssen alle Dokumente geöffnet werden und nacheinander in die Druckerwarteschlage geschickt werden. Das Splitten eines Dokumentes kann durch die Angabe der gewünschten Seitenzahlen erzielt werden, was jedoch auch nicht besonders intuitiv oder praktisch ist.

PDFArchitect ist ein weiteres Programm von pdfforge, welches eine Vielzahl an Modulen anbietet, welche einzeln für einen Aufpreis erworben werden können. Hierbei beispielsweise wird das Splitten und Mergen erleichtert, das Schwärzen von Informationen unterstützt oder OCR unterstützt.

Fazit: PDFCreator bietet sehr viele nützliche Funktionen für die Arbeit mit PDF-Dokumenten an. Es ist ein Open-Source Produkt und somit kostenlos nutzbar, bietet jedoch auf dieser Grundlage einige Einschränkungen wie das umständliche Splitten und Mergen. Optional

kann auf die kostenpflichtigen Zusatzmodule des PDFArchitect zugegriffen werden, die den Funktionsumfang um ein Vielfaches erhöhen. Eine weitere Alternative wäre die Kombination von PDFCreator mit einer weiteren freien Software, welche die fehlenden Funktionen ergänzt.

3.3.2.10 PDFtk

PDFtk Free ist eine Open-Source Software, die unter der GPL-Lizenz angeboten wird. Sie ist damit privat und in Unternehmen kostenlos nutzbar.

Auf der Website von PDFtk sind einige Hilfestellungen wie z.B. Anleitungen, Beispiele und ein Blog verfügbar, welche der Benutzer kostenfrei verwenden kann. Außerdem kann der Entwickler direkt per E-Mail kontaktiert werden, um fragen zu klären.

PDFtk Free ist für Windows XP, Vista, 7 und 8 verfügbar.

Da es sich hierbei um die „Free"-Version von PDFtk handelt, sind einige Funktionen nicht nutzbar. Um diese Funktionen freizuschalten muss die „Pro"-Version für $3,99 erworben werden. Die Free-Version umfasst das Zusammenfügen und Aufteilen von PDF-Dokumenten. Andere Funktionen wie das Hinzufügen von Wasserzeichen und die Verschlüsselung von Dokumenten sind nur in der Pro-Version verfügbar.

Neben der Free- und Pro-Version wird eine Server-Version angeboten, die einige fehlende Funktionen der Free-Version kostenfrei zur Verfügung stellt. Da es sich hierbei jedoch um ein kommandobasiertes Programm handelt, ist die Nutzung nicht intuitiv und sicherlich für einige Benutzer schwer verständlich.

Fazit: PDFtk bietet in der Pro-Version einige gute Funktionen, welche aber nicht kostenfrei zur Verfügung stehen. Die Free-Version ist ein gutes Split und Merge Programm und könnte in Kombination mit einem anderen Programm genutzt werden. PDFtk Server bietet gute Funktionen, welche jedoch schwer umzusetzen sind, wenn der Benutzer keine Kenntnisse mit kommandozeilenbasierten Programmen hat.

3.3.3 Auswertung der Testergebnisse

Dieses Kapitel zeigt drei Auswertungen der Softwarevergleichstests. Hierfür wurde zu Beginn ein Benchmarking der besten zwei PDF Lösungen mit dem dem aktuellen Produkt durchgeführt. Hier sollen sich im grafischen Vergleich Stärken und Schwächen der jeweiligen Produkte zeigen. Wie man in Abbildung 2 erkennen kann, sind jedoch die Testergebnisse ähnlich. In der Betrachtung der unfunktionellen Anforderungen gibt es kleine Unterschiede. Alle drei Lösungen laufen unter unterschiedlichen Lizenzen. Hierfür wurden die Punkte un-

terschiedlich vergeben. PDFCreator hat mit vier Punkten die höchste Punktzahl erzielt. Hier handelt es sich um ein Produkt unter der GPL Lizenz. PDF24Creator konnte nur 2 Punkte erzielen, da es sich bei der Lizenz um eine proprietäre Freeware handelt. Das Partnerunternehmen würde eine Open Source Lösung bevorzugen. Auf letztem Platz, abgeschlagen mit null Punkten, liegt das aktuell genutzte pdf Factory. Der Grund für die schlechte Bewertung begründet sich in der kostenpflichtigen proprietären Lizenz.

Im zweiten Teil werden funktionelle Anforderungen miteinander verglichen. Wie gut zu erkennen ist, unterscheiden sich die Produkte hier nur in vier Aspekten. Die Funktionen Split and Merge werden von den Prpgrammen PDFCreator und PDF24Creator nicht oder nur unzureichend angeboten. Aus diesem Grund kann hier die proprietäre Software gegenüber den freien Alternativen punkten. Die Anforderung der Zensur oder des Schwärzens von Informationen kann nur von der aktuellen Software pdf Factory angeboten werden. Die beiden anderen Anwendungssysteme bieten diese Funktion nicht an bzw. können nur über ein kostenpflichtiges Zusatzprogramm genutzt werden. Der finale Unterschied liegt bei der Unterstützung der Archivierungsfunktion PDF/A. PDF24Creator unterstützt dieses Archivierungsformat nicht.

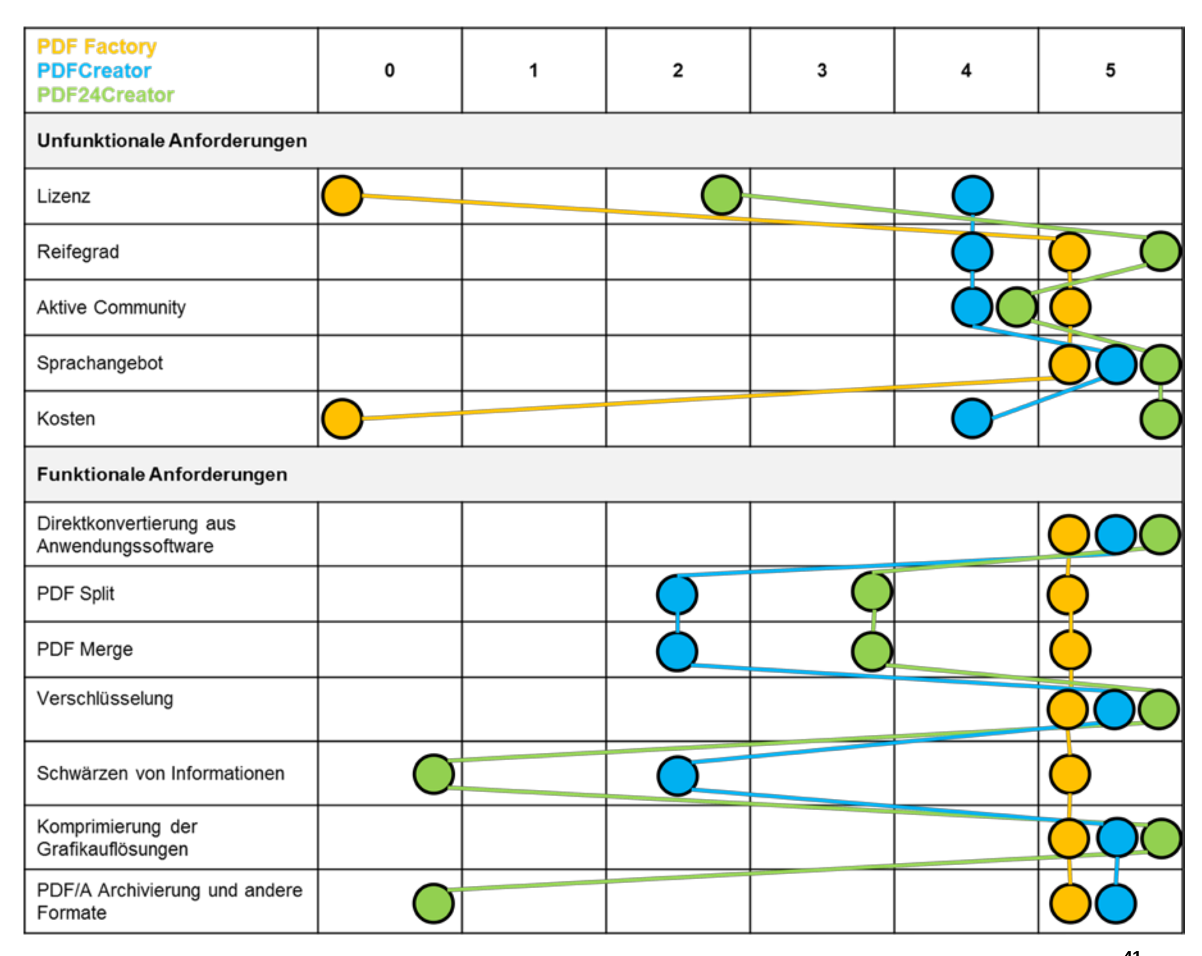


Abbildung 3: Benchmarking zwischen PDF Factory, PDF Creator und PDF24Creator[41]

[41] Eigene Darstellung

Vor Testdurchführung wurden Kriterien gesammelt. Hier wurden auch KO-Kriterien definiert, die unbedingt erfüllt werden müssen. Die Auswertung in Abbildung 3 gibt an, wie diese KO-Kriterien erfüllt wurden. Ausgegangen wird bei dieser Auswertung von einem maximalen Bewertungswert von fünf Punkten zur Bewertung. Diese fünf Punkte erhält ein Produkt, wenn es die Anforderung komplett erfüllt. Es kann jedoch auch zu Abzug wegen Nichterfüllung oder nur Teilerfüllung einer Anforderung kommen. Als schlechteste Bewertung, also bei kompletter Nichterfüllung der Anforderung gibt es null Punkte. In diese Auswertung sind nur die Ergebnisse der zehn Open Source und Freewarelösungen geflossen. PDF Factory wurde nicht berücksichtigt.

Im Schaubild ist deutlich zu erkennen, dass der Funktionsumfang der getesteten Software unterschiedlich ist. Nach der ersten Vorauswahl der Produkte wurden bereits mehrere Lösungen aussortiert. Der Hauptgrund lag hier in der mangelnden Offlinefunktion. Da dieser Punkt bereits in der Vorauswahl geprüft wurde, kann dieses KO-Kriterium mit 100% erfüllt werden. Dahingegen wird der Punkt der aktiven Community nur mit 50% erfüllt. Das soll nun nicht bedeuten, dass nur die Hälfte der Produkte eine aktive Community hat. Hier wurden auch die Umstände der Software mit in die Bewertung integriert. Ein Programm das zum Beispiel unter einer Freeware läuft und ein Forum hat, wurde nicht mit vollen fünf Punkten Bewertet. Hier wurden evtl. nur zwei oder drei Punkte vergeben. Wenn das letzte Update einer Software und die letzten Einträge im Forum hingegen innerhalb der vergangenen zwei Monate aufkommen, wurden hier vier oder 5 Punkte vergeben.

Die Anforderung, Textinhalte zu Schwärzen, bzw. zu zensieren wird von nur von einem Programm angeboten. Diese Funktion ist jedoch nicht kostenlos implementiert, sonder muss noch durch ein kostenpflichtiges Zusatzmodul erworben werden. Der Erfüllungsgrad dieser Anforderung ist sehr gering.

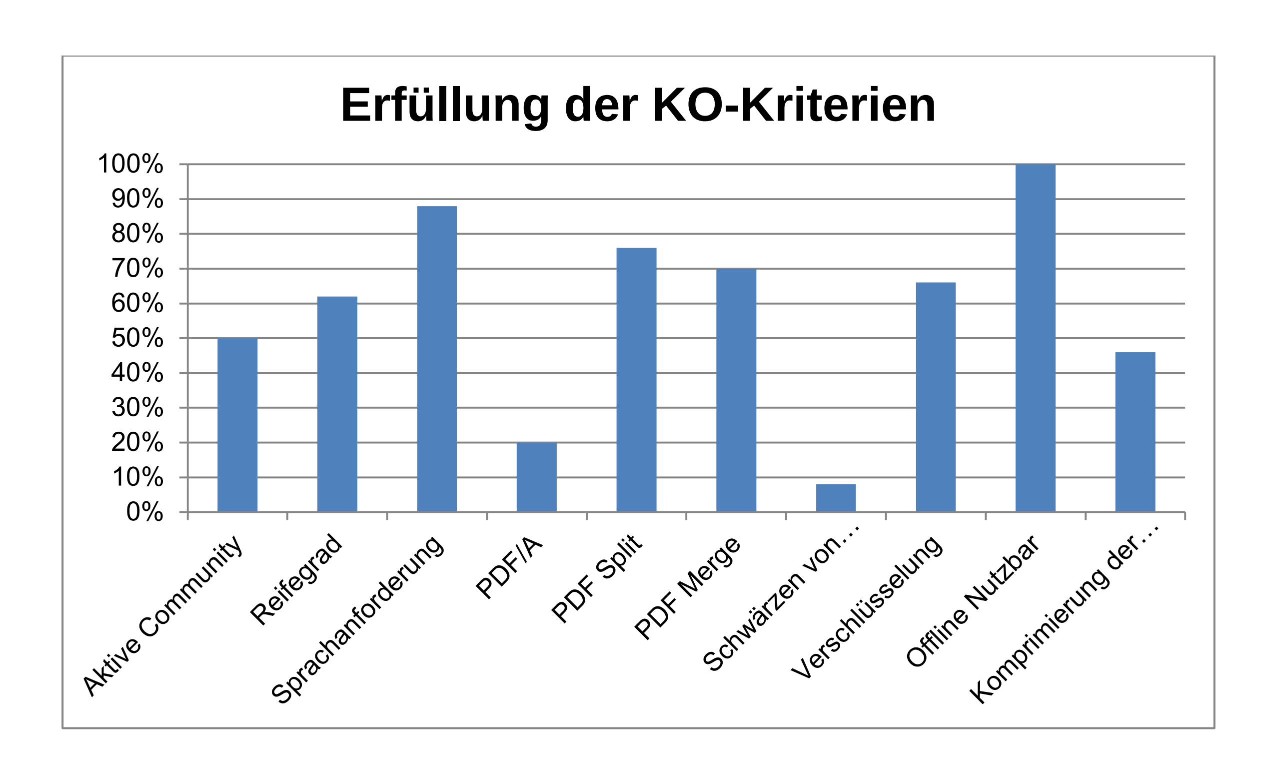


Abbildung 4: Erfüllung der KO-Kriterien[42]

In Abbildung 4 wird das Ergebnis der Softwaretests angezeigt. Hier ist zu erkennen, dass als Gewinner die Software PDFCreator ermittelt wurde. Fast ebenso gut ist die Software PDF24Creator. Die momentan genutzte Software pdf Factory hat in einer Vergleichsanalyse nur den Platz vier erreicht.

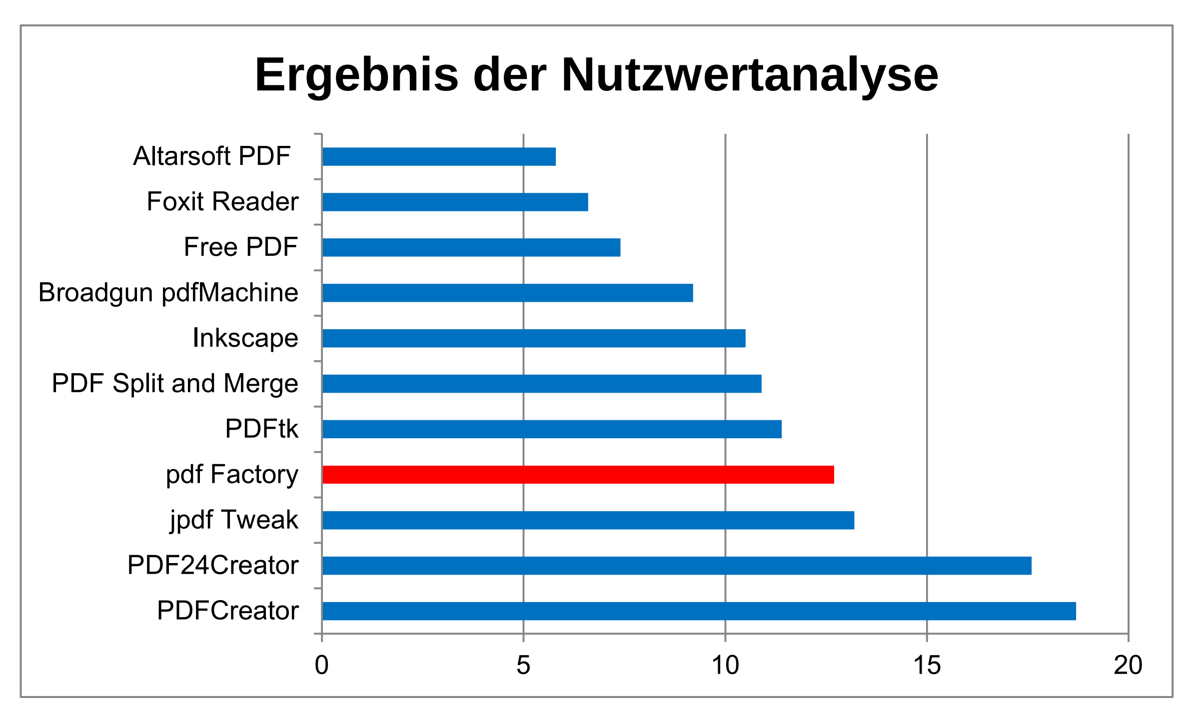


Abbildung 5: Ergebnis der Nutzwertanalyse[43]

[42] Eigene Darstellung

3.4 Monetärer Vergleich

Im monetären Vergleich wird das bestbewertete Produkt PDFCreator mit dem aktuell im Einsatz befindlichen pdfFactory Pro verglichen. Da das Versicherungsunternehmen mit den Herstellern von pdfFactory Pro ein Service-Level-Agreement (SLA) abgeschlossen hat, wird für den Kostenvergleich auf die seiten PDFCreator die gleichwertigste Variante Plus gewählt. In dieser Variante ist ein lebenslanger Support mit Updates enthalten. Bei der Kalkulation wird von einer Neueinführung des Produkts ausgegangen, da die genauen Kosten des SLA nicht bekannt sind.

Produkte	Kosten pro Lizenz	Kosten für 1000 Lizenzen
pdfFactory Pro	40,00 €	40.000,00 €
PDFCreator Plus	3,56 €	3.556 €
Gesamtersparnis	36,44 €	36.442 €

Tabelle 6: Monetärer Vergleich

Anhand der Rechnung wird ersichtlich, dass die Implementierung und der Umstieg auf PDFCreator Plus bei dem Versicherungsunternehmen zu einer beachtlichen Kostenersparnis von 36.442 € führen würde.

3.5 Ergebniszusammenfassung

Zu Beginn des Projektes wurden vom Partnerunternehmen verschiedene Anforderungen an ein alternatives Softwareprodukt gestellt. Diese Anforderungen wurden in einen Kriterienkatalog übernommen, welcher dann als Grundlage für die Software-Funktionstests möglicher Open-Source Produkte genutzt wurde.

Der Funktionsumfang der aktuell genutzten Anwendung PDF Factory ist sehr umfangreich. Ein Alternatives Produkt muss möglichst ebenso alle Funktionen anbieten, die PDF Factory anbietet. Da die Software PDF Factory als proprietäre Software im Unternehmen genutzt wird, sind hohe Lizenzkosten zu zahlen.

Für den Vergleichstest wurden zehn Anwendungen berücksichtigt. Die getesteten Produkte sind teilweise klassische Open Source Produkte. Jedoch wurden auch proprietäre Freeware mit in den Test integriert. Im Laufe des Tests, beziehungsweise nach der genauen Auswertung aller Softwaretests, gab es kein Alternativprodukt, das alle Anforderungen aus dem Katalog erfüllt. Als Lösung für diese Diskrepanz bietet sich daher eine Softwarekombination aus verschiedenen Produkten an.

[43] Eigene Darstellung

Die Sieger des Vergleichstests ist das Open Sourceprodukt PDFConverter und das Freewareprodukt PDF24Converter. Beide Anwendungen haben einen großen Funktionskatalog. Im Vergleichstest konnten diese zwei Produkte die Plätze eins und zwei belegen. Da beide Produkte die gewünschte Funktion Split and Merge (Zusammenfassen und Teilen) von PDF-Dokumenten nicht, beziehungsweise nur umständlich anbieten, soll hier ein Ergänzungsprodukt eingesetzt werden.

Der Vorteil dieser Kombinationslösung ist, dass alle Funktionen durch ein Open Source Produkte abgedeckt sind. Die Lizenzen für ein solches Produkt sind kostenlos und es besteht keine Zahlungsverpflichtung gegenüber der Entwicklercommunity. Falls neue Anforderungen entstehen, können diese Anforderungen an die Entwicklungscommunity des Open Source Produktes weitergeben werden. Eine Beteiligung an der zukünftigen Entwicklung des Produktes ist grundsätzlich möglich.

Sollte sich das Partnerunternehmen für eine Open Source Lösung entscheiden, werden Kosten für den Softwarewechsel auftreten. Eine Open Source Lizenz kostet selbst zwar nichts, jedoch entstehen genauso Kosten für Installation und Nutzung wie bei anderen Softwareprodukten. Ein weiterer Nachteil für den Wechsel zu einer Open Source Lösung ist, dass jede neue Software vor Einführung erst einmal im Unternehmensnetzwerk getestet werden muss. Ein solcher Test kann auch sehr aufwändig sein.

3.6 Livetest bei der .Versicherung

Zu Beginn des Projektes war ein Livetest beim betreuenden Partnerunternehmen vorgesehen. Bei der geplanten Durchführung des Livetestes ergaben sich jedoch Schwierigkeiten im Bezug auf die Umsetzung. Aus diesem Grund wurde der Livetest abgesagt. Zur Vorbereitung wurden jedoch bereits Testfälle und weitere Testdokumente geschrieben. Diese Daten werden in den kommenden Kapiteln beschrieben.

3.6.1 Testfälle

Um Festzustellen, ob ein Produkt den Anforderungen eines Unternehmens entspricht, sollte ein Produkt vorher in einer Testumgebung im Unternehmensnetzwerk getestet werden.

Für einen solchen Test wurden sechs beispielhafte Testfälle geschrieben. Diese Testfälle müssen mit den jeweiligen Unternehmenscomputern durchgeführt werden. Die Ergebnisse der jeweiligen Tests müssen festgehalten werden.

Da die sechs Testfälle nicht alle Funktionen der Software testen, müssen hier noch weitere Testfälle geschrieben werden. Diese Testfälle müssen speziell an das Unternehmen oder den Bereich angepasst werden.

3.6.1.1 Testfall #1

Nr.	Name	Beschreibung	Erwartetes Ergebnis
#1	PDF Dokument aus MS Word erstellen	Es soll aus einem MS Word ein PDF Dokument erstellt werden. Wenn das Software-Produkt als Drucker in der Geräteliste des Computers erscheint, soll hierfür dieser Drucker zum erstellen des PDF-Dokumentes genutzt werden. Vorgehen: 1. Gewünschtes MS Word Dokument aufrufen 2. MS Word Druckfunktion aufrufen 3. Dokument drucken mit PDF Software 4. PDF-Datei speichern 5. PDF-Datei prüfen	• Dokument wird als PDF Datei gespeichert • PDF Datei und Layout wird dargestellt wie das originale Word Dokument • PDF Datei wird fehlerfrei dargestellt

Tabelle 7: Testfall #1 – PDF Dokument aus MS Word erstellen

3.6.1.2 Testfall #2

Nr.	Name	Beschreibung	Erwartetes Ergebnis
#2	PDF Dokument aus MS Excel erstellen	Es soll aus einem MS Excel ein PDF Dokument erstellt werden. Wenn das Software-Produkt als Drucker in der Geräteliste des Computers erscheint, soll hierfür dieser Drucker zum erstellen des PDF-Dokumentes genutzt werden. Vorgehen: 1. Gewünschtes MS Excel Dokument aufrufen 2. MS Excel Druckfunktion aufrufen 3. Dokument drucken mit PDF Software 4. PDF-Datei speichern 5. PDF-Datei prüfen	• Dokument wird als PDF Datei gespeichert • PDF Datei und Layout wird dargestellt wie das originale Excel Dokument • PDF Datei wird fehlerfrei dargestellt

Tabelle 8: Testfall #2 – PDF Dokument aus MS Excel erstellen

3.6.1.3 Testfall #3

Nr.	Name	Beschreibung	Erwartetes Ergebnis
#3	PDF von einer beliebigen Webseite (z.B. www..Versicherung.de) mit Testtool erstellen	Es soll eine PDF Datei von einer Webseite, die im Browser aufgerufen wird,	• Dokument wird als PDF Datei gespeichert

		erzeugt werden. Wenn das Software-Produkt als Drucker in der Geräteliste des Computers erscheint, soll hierfür dieser Drucker zum erstellen des PDF-Dokumentes genutzt werden. Vorgehen: 1. Webseite in Browser aufrufen 2. Druckfunktion des Browsers aufrufen 3. PDF Software als Drucker auswählen und Drucken 4. PDF Datei speichern 5. PDF Datei prüfen Wichtig: Wenn im Unternehmen mehrere Browser genutzt werden, dann muss dieser Test mit jedem Browser durchgeführt werden.	• PDF Datei gibt die relevanten Inhalte der Webseite wieder • PDF Datei wird fehlerfrei dargestellt

Tabelle 9: Testfall #3 – PDF von einer beliebigen Webseite (z.B. www..Versicherung.de) mit Testtool erstellen

3.6.1.4 Testfall #4

Nr.	Name	Beschreibung	Erwartetes Ergebnis
#4	PDF aus einer verwendeten Anwendungssoftware heraus erstellen	In diesem Test soll die Möglichkeit getestet werden PDF Dokumente aus einer im Unternehmen genutzten Software heraus zu erstellen. Vorgehen: 1. Gewünschte Software und gewünschtes Dokument/Seite aufrufen 2. Druckfunktion aufrufen 3. PDF Software als Drucker auswählen und Drucken 4. PDF Datei speichern 5. PDF Datei prüfen Wichtig: • Für diesen Test bietet sich zum Beispiel ein Druck von einem erstellten Kundenangebot an • Um eine volle Funktionsfähigkeit zu testen, sollte dieser Testfall möglichst in jedem im Unternehmen genutzten	• Dokument wird als PDF Datei gespeichert • PDF Datei gibt die relevanten Inhalte der Webseite wieder • PDF Datei wird fehlerfrei dargestellt

		System durchgeführt werden.	

Tabelle 10: Testfall #4 – PDF aus einer ver-wendeten Anwen-dungssoftware heraus erstellen

3.6.1.5 Testfall #5

Nr.	Name	Beschreibung	Erwartetes Ergebnis
#5	PDF aus einem spe-ziellen .Versicherung-System heraus er-stellen	In diesem Test soll die Möglichkeit getes-tet werden PDF Dokumente aus einer speziell im Unternehmen genutzten Soft-ware heraus zu erstellen. Vorgehen: 1. Gewünschte Software und gewünsch-tes Dokument/Seite aufrufen 2. Druckfunktion aufrufen 3. PDF Software als Drucker auswählen und Drucken 4. PDF Datei speichern 5. PDF Datei prüfen Wichtig: • Für diesen Test bietet sich zum Bei-spiel ein Druck von einem erstellten Kundenangebot an • Um eine volle Funktionsfähigkeit zu testen, sollte dieser Testfall möglichst in jedem im Unternehmen genutzten System durchgeführt werden.	• Dokument wird als PDF Datei gespeichert • PDF Datei gibt die rele-vanten Inhalte der Web-seite wieder • PDF Datei wird fehlerfrei dargestellt

Tabelle 11: Testfall #5 – PDF aus einem spe-ziellen .Versicherung-System heraus erstellen

3.6.1.6 Testfall #6

Nr.	Name	Beschreibung	Erwartetes Ergebnis
#6	Angebot für einen Kunden als PDF erstellen (Split and Merge)	Dieser Testfall soll ein komplettes Ange-bot für einen Kunden speichern. Ein sol-ches Angebot soll aus mehreren ver-schiedenen Dokumenten heraus erstellt. Hierfür sollen die Funktionen Split and Merge genutzt werden. Vorgehen: 1. Dokumente zusammenstellen zum gemeinsamen Drucken. Folgende Beispiel Dokumente müssen dafür be-reit gestellt werden:	• Alle drei Dokumente Kundenangebot.pdf, AGB.pdf und Ange-bot.pdf wurden gespei-chert • Dokumente wurden in der gewünschten Rei-henfolge zusammenge-fügt • Seiten wurden wie ge-wünscht korrekt wieder

		• Angebot zum Vertrag • Aktuelle Allgemeine Geschäfts-bedingungen (AGB) • Versicherungs-Police • Unterschriebener Versicherungs-vertrag 2. PDF Software starten 3. Alle Dokumente in ein Dokument zusammenfassen und speichern unter Kundenangebot.pdf 4. Datei Kundenangebot.pdf aufrufen und Datei teilen in mehrere Dateien. Hierfür sollen einmal die AGBs separat als AGB.pdf gespeichert werden. Im weiteren Schritt sollen alle anderen Inhalte, ohne die AGBs als Angebot.pdf gespeichert werden. 5. Ergebnis prüfen	als separate Dokumente gespeichert • PDF Datei wird fehlerfrei dargestellt

Tabelle 12: Testfall #6 – Angebot für einen Kunden als PDF erstellen

3.6.2 Bewertungsprotokoll

Die Ergebnisse aus den vorherigen Testfällen müssen festgehalten und dokumentiert werden. Dafür kann folgendes Testprotokoll genutzt werden. In diesem Testprotokoll werden grundlegende Informationen zum Test dokumentiert. Diese Informationen sind zum Beispiel die Testplattform beziehungsweise das Betriebssystem auf dem die Software getestet wurde. Zusätzlich ist es relevant die jeweilige Softwareversion festzuhalten, auf der der Test beruht. Der Eintrag Tester und Datum des Tests dient ebenfalls zur Dokumentierung.

Programm	
Version	
Testplattform	
Datum	
Tester	
Zusammenfassung der Tests	

Tabelle 13: Metadaten des Testprotokolles

Die Dokumentation für jeden durchgeführten Testfall ist in vier unterschiedliche Bereiche Unterteilt. Im ersten Teil wird festgehalten, ob der jeweilige Test des Produktes durchgeführt wurde oder nicht. Im zweiten Schritt wird das Ergebnis festgehalten. Dieses bewertet direkt, ob das jeweilig getestete Produkt geeignet oder eher ungeeignet für das Unternehmen und die Nutzung ist. Es gibt noch den Punkt Bemerkung und Fazit. Hier können in Schriftform

noch weitere Informationen zum jeweiligen Testfall beschrieben werden. Diese Punkte sollten bei jedem Testfall möglichst ausgefüllt werden, da hiermit im Nachhinein das jeweilige Ergebnis nachvollzogen einfach werden kann.

Durchgeführt	☐ Ja	☐ Nein	☐ Abgebrochen
Ergebnis	☐ Geeignet	☐ Geeignet mit Einschränkungen	☐ Ungeeignet
Bemerkung			
Fazit			

Tabelle 14: Dokumentationsvorlage für einen Testfall

In den Allgemeinen Informationen des Bewertungsprotokolles wird noch der Punkt Zusammenfassung des Tests aufgeführt. Dieser soll ganz zum Schluss nach dem Durchführen aller Testfälle ausgefüllt werden. Die Zusammenfassung dient als Fazit des gesamten Testes und soll einem Leser die Möglichkeit geben, bereits zu Beginn des Dokumentes kurz und knapp einen Überblick und einen Eindruck über das getestete System zu gewähren.

Das Bewertungsprotokoll ist im Anhang 3 eingefügt.

3.6.3 Weitere Kriterien in einem Livetest

Die Kriterien der vorherigen Kapitel beziehen sich überwiegend auf funktionale und Produktnahe Kriterien. In einem Livetest in einem Unternehmensnetzwerk ist es noch sinnvoll, weitere Kriterien mit in Betracht zu ziehen. Ein Beispiel an zusätzlich relevanten Kriterien ist in Tabelle 15 aufgezeigt.

Kriterium	Beschreibung
Performance / Leistungsfähigkeit	Ein Programm sollte möglichst auf die Leistungsfähigkeit getestet werden. Dies bedeutet, dass zum Beispiel die Schnelligkeit der Informationsverarbeitung in einem Programm ein wichtiger Faktor ist. Im Falle der PDF-Verarbeitung macht es einen erheblichen Unterschied, ob die Erzeugung eines PDF-Dokumentes zehn Sekunden oder 2 Minuten dauert. Im Vergleich und im Einzelfall ist dieser Unterschied nicht viel, jedoch sind die Ausmaße für ein Großunternehmen enorm. Jede ungenutzte und unproduktive Minute an Arbeitszeit ist verschwendet und kostet ein Unternehmen Geld. Aus diesem Grund muss eine Softwarelösung möglichst performant laufen.
Systemanforderungen (Prozessor, Cache)	Das Partnerunternehmen arbeitet in Fachbereichen häufig mit einer Thin-Client-Serverlösung. Dies bedeutet, dass zum Beispiel kleine Computer mit sehr wenig eigener Rechenleistung aufgebaut werden. Diese Computer sind jedoch über das Netzwerk mit einem Server verbunden. Alle Daten und Programme werden der Nutzung einer Workstation bei jeder Anwendung direkt vom Server auf den jeweiligen Computer geladen. Eine solche Sitzung wird beendet und die Daten werden danach komplett von dem Thin Client gelöscht. Das Kriterium Systemanforderung und die benötigte Leistung ist daher ein relevanter Aspekt für ein Live-Test-Kriterium. Wenn die Konvertierung in ein PDF Dokument zentral über einen Server gesteuert wird, ist es natürlich wichtig, dass dafür so wenig wie möglich Rechenleistung verwendet wird. Hierzu muss natürlich auch die benötigte Speicherleistung berücksichtigt werden.
Installationsvarianten (.msi oder .exe)	Installationen im Partnerunternehmen laufen meistens zentral über das Netzwerk. Je nach Installationsmethode gibt es unterschiedliche Formate für die Installationsdatei. Diese Formate und die jeweiligen Vor- und Nachteile sollten in Betracht gezogen werden.

Tabelle 15: Livetest-Kriterien

4 Fazit und Handlungsempfehlung

Ziel dieser Arbeit war es eine Marktanalyse zu erstellen, die unterschiedliche Open Source Produkte zur PDF-Erzeugung und Bearbeitung miteinander und mit der momentan im Versicherungsunternehmen eingesetzten proprietären Software vergleicht. Dabei sollte das Ergebnis der Analyse eine Bewertungsgrundlage für eine Ablösung der kommerziellen Software und einer Implementierung von Open Source Produkten beim dem Versicherungsunternehmen darstellen.

Bereits bei der Recherche der Open Source Produkte fiel auf, das die auf dem Markt zur Verfügung stehende Anzahl an Open Source Software zur PDF-Erzeugung und Bearbeitung nicht ausreichend für eine aussagekräftige Marktanalyse sein würde. Aus diesem Grund wurden in Absprache mit dem Versicherungsunternehmen auch Freeware-Produkte zu der Analyse hinzugezogen.

Nach der Durchführung der Produkttests und der anschließenden Analyse der Testergebnisse stellte sich heraus, dass keines der getesteten Produkte alle KO-Kriterien des zuvor erstellten Kriterienkatalogs erfüllen konnte. Da der Kriterienkatalog jedoch speziell auf die Anforderungen des Versicherungsunternehmens zugeschnitten wurde, musste eine andere Lösung für die Problematik gefunden werden. Der im Folgenden beschriebene Ansatz stellt zwei alternative Konstellationen aus verschiedenen Produkten dar, welche kombiniert werden, um die Anforderungen des Versicherungsunternehmens erfüllen zu können.

Der Sieger des Produkttests war PDFCreator, eine Open Source Software, die mit einer Vielzahl an nützlichen Funktionen zur PDF-Erstellung und Bearbeitung überzeugen konnte. Das Programm konnte fast alle KO-Kriterien erfüllen und konnte sich dank der einfachen Bedienung und Funktionsvielfalt durchsetzen. Da die Split und Merge Funktion von PDFCreator relativ umständlich funktioniert und dies ein wichtiges KO-Kriterium für das Versicherungsunternehmen darstellt, soll im Rahmen dieser Handlungsempfehlung eine Kombination mit einem weiteren Open Source Produkt vorgestellt werden. Ein besonders guter Kandidat hierfür ist PDFSAM, ein Programm, das in den Tests zwar einige Anforderungen nicht erfüllen konnte, aber die benötigten Split und Merge Funktionen sehr gut durchführen kann. Diese Kombination aus PDFCreator und PDFSAM bietet eine Abdeckung nahezu aller wichtigen Funktionen und ist im Vergleich zu der proprietären Option komplett kostenfrei. Da es sich um Open Source Software handelt, fallen keinerlei Lizenzgebühren an, was ein klare Kostenersparnis für das Versicherungsunternehmen darstellt. Optional könnte die kostenpflichtige Plus-Version von PDFCreator genutzt werden,

welche lebenslangen Premiumsupport und Updates beinhaltet. Wie im monetären Vergleich gezeigt, ist die Kostenersparnis zwischen PDFCreator Plus und pdfFactory immens und bietet außerdem bei Bedarf auf Basis der GPL Möglichkeiten zur Anpassung des Sourcecodes.

Als Alternative dazu soll außerdem die Kombination aus PDF24 Creator und PDFSAM empfohlen werden. PDF24 Creator ist eine Freeware, die ebenfalls sehr gute Funktionen zur PDF-Erzeugung und Bearbeitung beinhaltet und zudem komplett kostenfrei erhältlich ist. Die Software hat einige Funktionen, wie beispielsweise eine integrierte Vorschau, die in manchen Bereichen einen Vorteil gegenüber PDFCreator darstellt. Auch die Split und Merge Funktion ist bei PDF24 Creator besser gelöst. Trotz dessen ist es zu empfehlen, PDFSAM als Split und Merge Programm zu installieren, da die Funktionen hier weitaus besser gelöst sind. PDF24 Creator bietet keine zusätzlichen Support-Optionen wie PDFCreator Plus an, verfügt jedoch über viele Kontaktstellen wie ein Forum, FAQ und einen ausführlichen Changelog.

Die beiden vorgeschlagenen Optionen bieten dem Versicherungsunternehmen eine solide Grundlage für die Bewertung der Frage, ob die proprietäre Software durch eine Open Source bzw. Kombination aus Freeware und Open Source Software ersetzt werden kann. Es konnte klar herausgearbeitet werden, dass mit Hilfe der gegebenen Handlungsvorschläge große Kostenaufwendungen eingespart werden können. Außerdem sind sowohl bei der Open Source als auch Freeware Lösung konstante Updates enthalten, die ein nachhaltiges Arbeiten mit der Software ermöglichen. Im Vergleich dazu muss für Updates einer proprietären Lizenz bezahlt werden, was die Kosten weiter erhöht. Des Weiteren wird die mögliche Abhängigkeit von dem Hersteller unterbunden, da besonders die reine Open Source Lösung offen für Veränderungen ist und Verknüpfungen mit diversen Schnittstellen erlaubt.

Schlussendlich muss sich das Versicherungsunternehmen für eine der Option entscheiden. Die Gefahr einer ungewissen Weiterentwicklung und teilweise keinen direkten Support durch den Entwickler kann äußerst abschreckend wirken und dazu führen, dass die Entscheidung auf die sicherere Variante der proprietären Lösung fällt. Die Möglichkeiten, die Open Source Software bietet, sei es nun monetär oder funktional, sollten jedoch bei der Entscheidungsfindung besonders in Betracht gezogen werden.

5 Anhang

Anhangverzeichnis

Anhang 1 – Vorlage des Kriterienkataloges mit Gewichtungen

Kriterien	KO-Kritierum	Gewichtung		Produkt		
Unfunktionale Anforderungen				**Begründung**	**Bewertung**	**Punktzahl**
Generell		**20%**				0,0
Hersteller		0				
Website		0		-		
Lizenz		5				
Aktive Community (Letztes Update)	x	5				
Reifegrad	x	5				
Systemanforderung (mindestens Windows 7)	x	5				
Sprache (mindestens Deutsch und Englisch)	x	5				
Umfang (Konverter, Editor, etc.)		3				
Technische Basis (Ghostscript, etc.)		3				
PDF-Spezifikation (unterstützte PDF-Formate)		1				
Kosten		**20%**				0,0
Kosten für Produkt		5				
Folgekosten für Support		5				
Total Unfunktionale Anforderungen		**40%**				**0**

Funktionale Anforderungen			Begründung	Bewertung	Punktzahl
Funktionalität		**40%**			**0,0**
PDF/A	x	5			
PDF/UA		3			
PDF/X		3			
Kompatibilität mit anderen Programmen (Add-ins)		1			
Direkt-Konvertierung Microsoft Office 2010		2			
Direkt-Konvertierung OpenOffice		2			
Personalisierung (Serienbrief, etc.) Autonmatisches Ausführen aus einer Datenbank		2			
PDFs zusammenfügen (merge)	x	5			
PDFs aufteilen (split)	x	5			
Briefpapier / Wasserzeichen		3			
Schwärzen von Informationen	x	5			
Verschlüsselung	x	5			
Zugriffsschutz		4			
Digitale Signatur		3			
Automatisierte Vorgänge		1			
OCR		1			
Offline nutzbar	x	5			
Usability		**20%**			**0,0**
Intuitive Bedienung		4			
Direkter E-Mail-Versand		2			
Vorschau		1			
Benutzerhilfe		3			
Dokumentation		3			
Druckvorlagen		2			
Maximale Dokumentenauflösung (in dpi)		2			
Komprimierung der Grafikauflösung	x	5			
Total Funktionale Anforderungen		**60%**			**0,0**

Anhang 2 – Systemanforderungen der durchgeführten Tests

Altarsoft PDF Converter

Testplattform	Windows 7 Enterprise
Software-Webseite	www.altarsoft.com
Getestete Version	PDF Converter 1.1 PDF Reader 1.2 Split Files 1.72

Broadgun pdfMachine

Testplattform	Windows 7 Enterprise
Software-Webseite	www.broadgun.de
Getestete Version	14.73

Foxit Reader

Testplattform	Windows 7 Enterprise
Software-Webseite	www.foxitsoftware.com
Getestete Version	6.0.4.0719

Free PDF

Testplattform	Windows 7 Enterprise
Software-Webseite	freepdfxp.de
Getestete Version	Ghostscript 9.07 FreePDF 4.14

Inkscape

Testplattform	Windows 7 Enterprise
Software-Webseite	inkscape.org
Getestete Version	0.91, 32bit-Windows EXE-Installer

Jpdftweak

Testplattform	Windows 7 Enterprise
Software-Webseite	jpdftweak.sourceforge.net
Getestete Version	1.1, Binary download, compact version

PDF Split and Merge

Testplattform	Windows 7 Enterprise
Software-Webseite	http://www.pdfsam.org
Getestete Version	PDFsam Basic 2.2.4

PDF24 Creator

Testplattform	Windows 7 Enterprise
Software-Webseite	de.pdf24.org
Getestete Version	6.9.2

PDFCreator

Testplattform	Windows 7 Enterprise
Software-Webseite	de.pdfforge.org
Getestete Version	2.0.2

pdftk

Testplattform	Windows 7 Enterprise
Software-Webseite	pdflabs.com
Getestete Version	2.0.2

Anhang 3 – Bewertungsprotokoll

Bewertungsprotokoll

Programm	
Version	
Testplattform	
Datum	
Tester	
Zusammenfassung der Tests	

Testfall #1

Durchgeführt	☐ Ja	☐ Nein	☐ Abgebrochen
Ergebnis	☐ Geeignet	☐ Geeignet mit Einschränkungen	☐ Ungeeignet
Bemerkung			
Fazit			

Testfall #2

Durchgeführt	☐ Ja	☐ Nein	☐ Abgebrochen
Ergebnis	☐ Geeignet	☐ Geeignet mit Einschränkungen	☐ Ungeeignet
Bemerkung			
Fazit			

Testfall #3

Durchgeführt	☐ Ja	☐ Nein	☐ Abgebrochen
Ergebnis	☐ Geeignet	☐ Geeignet mit Einschränkungen	☐ Ungeeignet
Bemerkung			
Fazit			

Testfall #4

Durchgeführt	☐ Ja	☐ Nein	☐ Abgebrochen
Ergebnis	☐ Geeignet	☐ Geeignet mit Einschränkungen	☐ Ungeeignet
Bemerkung			
Fazit			

Testfall #5

Durchgeführt	☐ Ja	☐ Nein	☐ Abgebrochen
Ergebnis	☐ Geeignet	☐ Geeignet mit Einschränkungen	☐ Ungeeignet
Bemerkung			
Fazit			

Testfall #6

Durchgeführt	☐ Ja	☐ Nein	☐ Abgebrochen
Ergebnis	☐ Geeignet	☐ Geeignet mit Einschränkungen	☐ Ungeeignet
Bemerkung			
Fazit			

Anhang 4 – Vorauswahl an möglicher Software

Software	Lizenz	Betriebssystem
Altarsoft PDF Converter	Freeware	Windows
BeCyPDFMetaEdit	Freeware	Windows
Broadgun pdfMachine	Freeware	Windows
Foxit Reader	Freeware	verschiedene
FreePDF	Freeware	Windows
Inkscape	GPL	verschiedene
jpdftweak	AGPL	verschiedene
NaivPDF	Freeware	verschiedene
Online-Umwandeln.de	Freeware	verschiedene
OpenOffice	GPL	verschiedene
PDF Chain	GPL	Linux
PDF Mergy	Freeware	verschiedene
PDF Split and Merge	GPL	verschiedene
PDF24 Creator	Freeware	Windows
PDF2TXT	Freeware	verschiedene
PDFCreator	GPL	Windows
PDFedit	GPL	verschiedene
pdftk	GPL	verschiedene

6 Quellenverzeichnis

Selbstständige Bücher und Schriften

Grimm, R. / Schuller, M. / Wilhelmer, R. (2014): Portfoliomanagement in Unternehmen, Leitfaden für Manager und Investoren, 1. Aufl., Wiesbaden: Springer Fachmedien

Verzeichnis der Internet- und Intranetquellen

Adobe (2007): JavaScript for Acrobat API Reference, Version 8.1, http://wwwimages.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/js_api_reference.pdf, Abruf: 03.02.2015

Adobe (2008): Adobe Supplement to the ISO 32000, Version 1.7, http://wwwimages.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/adobe_supplement_iso32000.pdf, Abruf: 03.02.2015

Bienz, T./Cohn, R./Meehan, J.R. (1996): Portable Document Reference Manual, Version 1.2, http://www.jbw.pl/firemka/pdfref12.pdf, Abruf: 03.02.2015

BITKOM (2006): Open Source Software – Rechtliche Grundlagen und Hinweise, http://www.bitkom.org/files/documents/BITKOM_Publikation_OSS_Version_1.0.pdf, Abruf: 03.02.2015

Bund.de (o.J): Vom Lastenheft zum Kriterienkatalog, http://gsb.download.bva.bund.de/BIT/V-Modell_XT_Bund/V-Modell%20XT%20Bund%20HTML/e91e12541a875f7.html, Abruf: 02.02.2015

Diedrich, O. (2009): Trendstudie Open Source, http://heise.de/-221696, Abruf: 03.02.2015

Drümmer, O. (2011): PDF/VT im Kontext von PDF/X, PDF/A, PDF/UA, http://www.pdfa.org/wp-content/uploads/2011/11/PDF-VT-im-Kontext-von-PDF-X-PDF-A-und-PDF-UA.pdf, Abruf: 03.02.2015

Eggeling, T. (2008): Ausfüllbare PDF-Formulare erstellen, http://www.pcwelt.de/tipps/Open-Office-Ausfuellbare-PDF-Formulare-erstellen-1243625.html, Abruf: 03.02.2015

Heiermann, C. (2013): Sicher archivieren mit PDF/A, http://www.computerwoche.de/a/sicher-archivieren-mit-pdf-a,2530759, Abruf: 03.02.2015

Heinrich, H. / Holl, F. / Menzel, K. / Mühlberg, J. / Schäfer, I / Schüngel, H. (2006): Metastudie – Open-Source-Software und ihre Bedeutung für Innovatives Handeln, http://www.bmbf.de/pubRD/oss_studie.pdf, Abruf: 03.02.2015

Jaeger, T. / Schulz, C (2005): Gutachten zu ausgewählten rechtlichen Aspekten der Open Source Software, http://www.ifross.org/ifross_html/art47.pdf, Abruf: 03.02.2015

Jelitto, M. (2002): Methode: Kriterienkatalog, http://www.evaluieren.de/evaluat.ion/methoden/0001.htm, Abruf: 02.02.2015

Open Source Initiative. (o.J): The Open Source Definition, http://opensource.org/osd, Abruf: 03.02.2015

Renner, T. / Vetter, M. / Rex, S. / Kett, H. (2005): Open Source Software: Einsatzpotenziale und Wirtschaftlichkeit – Eine Studie der Fraunhofer-Gesellschaft, http://wiki.iao.fraunhofer.de/images/6/63/Fraunhofer-Studie-Open-Source-Software.pdf, Abruf: 03.02.2015

Röder, H./Franke, S./Müller, C./Przybylski, D. (o.J.): Ein Kriterienkatalog zur Bewertung von Anforderungsspezifikationen; http://pi.informatik.uni-siegen.de/stt/29_4/03_Technische_Beitraege/Roeder-STT-BewertungAnfSpez.pdf, Abruf: 02.02.2015

Schubert, T. (2002): PDF Workflow in der Druckvorstufe: Ansätze zur automatisierten Produktion, https://books.google.de/books?id=HehjAQAAQBAJ&printsec=frontcover&hl=de#v=onepage&q&f=false, Abruf: 03.02.2015

Weber, M. (2008): Seminar „E-Learning“: Kriterienkataloge, https://ddi.informatik.uni-erlangen.de/teaching/SS2008/SeminarE-Learning/seminar-e-learning-ss2008-weber-vortrag.pdf, Abruf: 01.02.2015

Prepressure (2013): The history of PDF, http://www.prepressure.com/pdf/basics/history, Abruf: 03.02.2015



Auswahl und Bewertung von Open Source Schnittstellentransformationstools

Schriftliche Ausarbeitung
im Rahmen der Lehrveranstaltung „Integrationsseminar“
für das Kompetenzzentrum Open Source (KOS)

Vorgelegt von

Isabelle Pfahler, Isabelle Schwarz
Andreas Bucher, Patrick Espenschied
Sebastian Ober, Yannick-Tjard Schuetz

am 04.02.2015

Fakultät Wirtschaft
Studiengang Wirtschaftsinformatik
WWI 2012 E

Inhaltsverzeichnis

Abkürzungsverzeichnis

AHP	Analytic Hierarchy Process
BSD	Berkeley Software Distribution
C.I.	Consistency Index
C.R.	Consistency Ratio
CSV	Comma Separated Value
DACH	Bezeichnung für die Länder Deutschland, Österreich, und Schweiz
DTD	Document Type Definition
ECMA	European Computer Manufacturers Association
EDI	Electronic Data Interchange
EDIFACT	Electronic Data Interchange for Administration, Commerce and Transport
ETL	Extract-Transform-Load
GDV	Gesamtverband der Deutschen Versicherungswirtschaft e. V.
GPL	GNU General Public License
HTML	HyperText Markup Language
http/ https	Hypertext Transfer Protocol/ Hypertext Transfer Protocol Secure
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
IMAP	Internet Message Access Protocol
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
MIT	Massachusetts Institute of Technology
OS	Operating System oder Open Source
OSL	Open Software License
OSI	Open Source Initiative
OSS	Open Source Software
POP3	Post Office Protocol 3
R.I.	Random Index
SGML	Standard Generalized Markup Language
SMTP	Simple Mail Transfer Protocol
TDCC	Transportation Data Coordinating Committee
UN/EDIFACT	United Nations Directory for EDIFACT
VU	Versicherungsunternehmen
W3C	World Wide Web Consortium
XML	Extensible Markup Language
XML-RPC	Extensible Markup Language - Remote Procedure Call

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Problemstellung

Die .Versicherung hat die Schwierigkeit, dass die Kommunikation und der Informationsaustausch mit externen Partnern und staatlichen Einrichtungen in den letzten Jahren deutlich zugenommen haben. Die unterschiedlichen Partner und Einrichtungen benutzten in ihren Systemen zum Teil andere Datenformate als die .Versicherung. Daher müssen Datenformate oft in andere Formate transformiert werden. Hierfür stehen derzeit nur Lösungen bzw. Anwendungen zur Verfügung, die entweder nur einzelne Datenformate transformieren oder die anfallende Datenmenge nicht verarbeiten können. Deshalb sucht die .Versicherung nach einer Open Source-Software, welche eine Lösung für das beschriebene Problem liefert.

1.2 Zielsetzung

Ziel der Arbeit ist es, anhand von definierten Kriterien eine Auswahl an drei bis fünf Open Source Lösungen zur Datentransformation zu treffen. Die zwei besten Lösungen sollen in einem Prototyp umgesetzt und getestet werden.

1.3 Methodisches Vorgehen

Zu Beginn dieser Arbeit wird Open Source Software und deren Lizenzmodelle vorgestellt, sowie der Begriff der Datentransformation definiert. Anschließend werden in einer Übersicht sämtliche Datenformate, die im Zuge dieser Arbeit berücksichtigt werden müssen, dargelegt. Der Hauptteil dieser Arbeit orientiert sich im Aufbau an den ersten Schritten eines Proof of Concept (PoC). Typischerweise werden zu Beginn eines PoC die Erfolgskriterien bestimmt. Dies erfolgt in dieser Arbeit in Form eines Kriterienkatalogs, der sowohl technische als auch wirtschaftliche Kriterien auflistet, die mit dem Auftragsgeber abgestimmt werden. Da keine eigene Software programmiert, sondern eine passende Open Source Lösung gefunden werden soll, besteht der zweite Schritt aus einer Analyse aller auf dem Markt verfügbaren Open Source Lösungen zur Datentransformation. Hierzu wird eine Nutzwertanalyse bzw. AHP-Analyse durchgeführt. Ziel der Analyse ist es, die Auswahl an Produkten auf drei bis fünf einzuschränken. Im dritten Schritt dieses PoC, wird zu den, laut Analyse, zwei besten Produkten jeweils ein Prototyp erstellt. Diese Prototypen werden dann umfassend auf, durch

den Auftraggeber definierte Parameter, getestet. Mit der Evaluation der Tests ist diese Arbeit abgeschlossen. Weitere Schritte eines PoC werden somit in dieser Arbeit nicht betrachtet.

2 Definitionen

Um die im weiteren Verlauf der Arbeit verwendeten Begriffe zum Thema Open Source und Lizenzen vor angemessenem Hintergrund verwenden zu können, wird in diesem Kapitel näher darauf eingegangen. Besonders im Hinblick auf das Thema der gesamten Arbeit ist es wichtig, Open Source näher zu definieren.

2.1 Open Source

1985 wurde erstmals der Begriff „free software“ von der Free Software Foundation definiert, heutzutage besser bekannt als Open Source Software (OSS). Diese Bezeichnung entstand 1998 mit der Gründung der Open Source Initiative (OSI), die sich für die Förderung von Open Source Software einsetzt. Open Source prägt seit Jahren die Software-Industrie und wird immer wichtiger – auch für Unternehmen.[1]

Das bedeutendste Merkmal von Open Source Software ist, dass der Quellcode frei zugänglich und für jeden Interessenten einsehbar ist. Zusätzlich ist es jedem erlaubt, diesen Code selbst zu nutzen, weiterzuentwickeln und zu verbessern. Dies geschieht meist zusammen in den jeweiligen Communities der Projekte. Bei der Verbreitung eines weiterentwickelten Werkes sind allerdings die verschiedenen Open Source Lizenzen zu beachten. Charakteristische Hauptmerkmale von Open Source Software sind die „Lizenz der Software [,] nicht-kommerzielle Einstellungen [, ein] hoher Grad an Kollaboration bei der Programmentwicklung [und eine] starke räumliche Verteilung der Entwickler“[2]. Nicht zu verwechseln ist die Open Source Software mit sogenannter „Freeware“. Diese ist zwar kostenlos, jedoch erfüllt sie nicht zwingend auch das wesentlichste Merkmal von Open Source Software: der frei zugängliche Quellcode.[3]

[1] Vgl. Ueda, M. (2005), S. 381 f.
[2] Nüttgens, M. (2014)
[3] Vgl. Nüttgens, M. (2014)

2.2 OS Lizenzen

Um den Code trotz der vielen Freiheiten zu kontrollieren, gibt es inzwischen weit über 50 verschiedene Lizenzen, welche von der OSI auf Open Source – Kompatibilität geprüft und zugelassen wurden. Hier unterscheidet man vor allen Dingen zwischen drei Arten: „free-for-all“, „keep-open“ und „share-alike“. Unter ersterem versteht man eine Lizenz, die nur verlangt, dass bei Änderungen oder Erweiterungen des ursprünglichen Codes der Erstautor akkreditiert wird. Dies bedeutet, dass eine OSS unter dieser Lizenz problemlos von Dritten verbessert und anschließend von denselben urheberrechtlich geschützt werden kann. Dadurch wird die Software proprietär, der Source Code ist also nicht mehr einseh- oder veränderbar von Dritten. Der Autor der Originalsoftware tritt fast alle Rechte ab. Beispiele einer solchen Lizenz sind BSD (Berkely Software Distribution), MIT Lizenzen und auch Apache. „Keep-Open“ dagegen bedeutet, dass Veränderungen immer die gleiche Lizenz tragen müssen wie das ursprüngliche Produkt; die Software bleibt also Open Source. Eine Ausnahme gibt es hier bei sehr großen Lösungen, welche zu einem geringen Teil auf einer OSS mit „keep-open“ Lizenz aufbaut. Hier haben Dritte das Recht, ihr Programm urheberrechtlich zu schützen, mit Softwarepatenten zu versehen, oder sie auf andere Art und Weise proprietär zu machen. Die LGPL (GNU Lesser General Public License) oder die Mozilla Public License sind klassische Beispiele dieser Lizenzart. Zuletzt sind Lizenzen wie die GPL (GNU General Public License) und OSL (Open Software License) zu erwähnen, welche als „share – alike“ oder Copyleft bezeichnet werden. Werke, die aus OSS mit dieser Art von Lizenz hervorgehen, müssen auch unter derselben weiterverbreitet werden. Somit wird sichergestellt, dass die Software und die daraus resultierenden Werke weiterhin Open Source bleiben.[4]

Im folgenden Verlauf des Kapitels wird genauer auf die Apache 2.0 und GPLv3 eingegangen, da erstere von Talend und letztere von Bots als Lizenz eingesetzt wird.

2.2.1 Apache 2.0

Seit 2004 besteht die Apache Version 2.0. Wie bereits erwähnt handelt es sich bei dieser Lizenz um eine „free-for-all“-Lizenz. Grundsätzlich bedeutet das, dass jegliche Werke unter dieser Lizenz royalty-free sind. Nutzer werden also in keinerlei Weise eingeschränkt, was die Verwendung der Software anbelangt und können sie nach eigenem Ermessen modifizieren, weitergeben, verkaufen, mit neuen Rechten und Patenten versehen, etc. Wenn der Source Code allerdings gegen bestehende Patente oder Gesetze verstößt, so werden jeglicher modifizierten Software dieser Teil sowie die daran geknüpften Rechte entzogen. Grundsätzlich

[4] Vgl. Engelfriet, A. (2010), S. 48 f.

jedoch sind bei der Weitergabe von Source Code unter Apache 2.0 folgende Dinge zu beachten:

- Jede OSS die Teile dieses Source Codes verwendet muss mit dem Produkt eine Kopie Apache 2.0 Lizenz liefern
- Jeglichen Modifikationen müssen deutlich gekennzeichnet werden
- Übernommene Source Code Teile müssen als Apache 2.0 gekennzeichnet werden, das gesamte Werk bzw. die Modifikationen können mit einer veränderten oder gänzlich verschiedenen Lizenz versehen werden
- Angehängte Notizen zur Lizenz oder Software müssen weitergeleitet werden (es sei denn sie betreffen keinen Teil des neuen Programms) und können erweitert werden, insofern die Anmerkungen nicht in die Lizenzregelungen eingreifen

Autoren haften nur beschränkt und geben keinerlei Garantie, wenn es nicht rechtlich notwendig ist. Zum weiteren Schutz der Lizenzgeber haben Nutzer keinerlei Anrecht auf deren Trademarks (Namen, Logo, etc.). Sobald das Produkt als proprietäre Software weitergegeben wird haftet, wenn es so in den Lizenzbestimmungen festgelegt wurde, der Anbieter, niemals jedoch der Verfasser des Source Codes, auf dem das Werk basiert.[5]

2.2.2 GPLv3

Die dritte Version der GNU General Public License wurde am 09. Juli 2007 veröffentlicht. Die Free Software Foundation hat zu diesem Zeitpunkt bereits verschiedene Lizenzmodelle veröffentlich, wie beispielsweise die GPL Version 1.0 und 2.0. In der neuesten Version gelten vier Grundsätze zur Freiheit: jeder darf Werke unter GPLv3 weiterverbreiten, auch für Geld. Man hat das Anrecht darauf den Source Code zu erhalten, sowie den gesamten Code oder Teile davon für seine eigenen Projekte zu nutzen und zu modifizieren, insofern diese ebenfalls unter GPLv3 laufen. Zuletzt besitzt jeder Nutzer das Recht über diese Rechte informiert zu werden. Um sicher zu stellen, dass diese Rechte niemandem entzogen werden, der an der Software mitwirkt oder mitgewirkt hat, hat der Distributor bestimmte Pflichten. Er ist dafür verantwortlich, dass alle Nutzer die gleichen Rechte an dem veränderten Werk haben, wie er selbst am Original und dass sie auch über diese Rechte Bescheid wissen. GPLv3 Software muss also in jedem Fall (ob kostenpflichtig oder nicht) unter GPLv3 weitergegeben werden, da es sich um eine „share–alike“ Lizenz handelt. Um die Autoren zu schützen sieht die

[5] Vgl. Apache Software Foundation (2015)

GPLv3 vor, dass diese keinerlei Garantie für modifizierte Software o.Ä. übernehmen und nur beschränkt haftbar sind. Somit ist gewährleistet, dass sie nicht für Fehler Dritter belangt werden können. Ebenfalls legt GPLv3 fest, dass jegliche unter dieser Lizenz geführten Produkte nicht durch neue Patente proprietär gemacht werden können.[6]

Grundsätzlich sind bei der Verbreitung von Produkten, die GPLv3 lizensiert sind folgende Dinge zu beachten[7]:

- Die modifizierte Arbeit muss klar gekennzeichnet werden
- Es muss deutlich ausgewiesen werden, dass das Werk GPLv3 lizensiert ist
- Jegliche Modifikation muss ebenfalls unter derselben Lizenz verbreitet werden, damit gewährleistet wird, dass alle Open Source Rechte intakt bleiben
- Sobald das Produkt eine interaktive Oberfläche hat, müssen sich auf jeder Seite entsprechende rechtliche Hinweise befinden

2.2.3 Gegenüberstellung

Aus vorangegangenen Kapitel werden einige Unterschiede zwischen Apache 2.0 und GPLv3 ersichtlich. Im Folgenden wird beschrieben, welche Vor- und Nachteile sich daraus ergeben.

Der bedeutendste Unterschied zwischen den beiden Lizenzen ihr die Lizenzart. Während Apache 2.0 „free-for-all“ ist und somit die Lizensierung bei der Distribution komplett geändert werden kann, so ist die GPLv3 als „share – alike“ Lizenz hauptsächlich auf den Erhalt der Open Source Grundsätze bedacht, die den Vertrieb proprietärer Software ablehnen.

Der Vorteil einer Copyleft-Lizenz ist, dass es Distributoren, wie bereits erwähnt wurde, nicht möglich ist, auf der Grundlage dieser OSS ein proprietäres Produkt zu entwickeln. Dies bedeutet, die Autoren des Source Codes genießen die Sicherheit, dass niemand ihr Produkt für rein kommerzielle Zwecke missbraucht, sondern in erster Linie den Code zum Allgemeinwohl aller verbessern möchte. Somit liegt hier eine Win-Win Situation vor: der Autor profitiert von den Beiträgen Dritter, welche wiederum einen Vorteil durch die Arbeit des Autors gewinnen. Auf der anderen Seite kann diese Lizenzart abschreckend wirken und somit potentielle Verbesserungen zeitlich verzögern. Jegliche Software muss nach Einbindung von GPLv3 Source Code ebenfalls unter GPLv3 lizensiert sein und wird somit Open Source. Dies kann

[6] Vgl. Free Software Foundation (2007)
[7] Vgl. Free Software Foundation (2007)

Entwickler gänzlich von der Teilnahme am Projekt abhalten, da sie nicht all ihr Gedankengut allgemein verfügbar machen möchten.

Aber auch „free-for-all“ – Lizenzen müssen mit einem kritischen Auge betrachtet werden. Da jede modifizierte Software nach den Wünschen des Distributors lizensiert werden kann, ist es Dritten möglich eine verbesserte Version des Produkts proprietär zu vertreiben und somit der Allgemeinheit alle Open Source Rechte zu entziehen. Der Nutzer hat also in diesem Fall einen Vorteil gegenüber dem Autor. Andererseits laden Apache 2.0 Projekte stark zur Mitarbeit ein, da Entwickler ihren Code schützen können bzw. auch ein kommerzielles Nutzen des Ergebnisses ihrer Arbeit möglich ist.

Abschließend lässt sich sagen, dass die Verwendung der Lizenz von der Absicht des Autors und der Software abhängig gemacht werden sollte. Apache 2.0 lädt zur Mitarbeit ein und begünstigt somit tendenziell eine schnellere Verbreitung und Etablierung der Software, bietet jedoch im Gegenzug hauptsächlich dem Nutzer und nicht dem Autor Vorteile. GPLv3 dagegen stellt sicher, dass der Autor und alle Nutzer durch die Modifikationen Dritter einen Mehrwert erlangen, kann aber potentielle Entwickler abschrecken.[8]

2.3 Datentransformation

Es gibt viele verschiedene Möglichkeiten Daten, darzustellen und zu speichern. Dabei können für ein und denselben Datensatz verschiedene Datenformate verwendet werden. Häufig ist eine Transformation von Datensätzen von ihrem derzeitigen zu einem anderen Datenformat notwendig, um sie für andere Systeme lesbar zu machen. Diesen Vorgang bezeichnet man als Datentransformation oder auch Datenkonvertierung.

Um eine Datentransformation zu vollziehen wird ein Datenkonverter, also eine spezielle Software zur Überführung der Daten in das neue Format, benötigt.

Bei der Konvertierung unterscheidet man zwischen verlustfreier, verlustbehafteter und sinnhafter Konvertierung. Von einer verlustfreien Konvertierung spricht man, wenn die neue Datei alle Informationen des Originals beinhaltet und bis auf den Datentyp unverändert ist. Verlustbehaftete Konvertierungen führen dazu, dass bestimmte Informationen in der neuen Datei fehlen können oder anders dargestellt werden, da sie im neuen Datenformat entweder nicht wiedergegeben werden können oder verlusthaft komprimiert wurden. Sinnhafte Konvertierungen übertragen alle Informationen, die als wesentlich angesehen werden. Das kann mit

[8] Vgl. Engelfriet, A. (2010), S. 48 f.

oder ohne Informationsverlust geschehen. Außerdem können die neuen Dateien mit zusätzlichen Informationen angereichert werden.

Eine Umwandlung in andere Datenformate kann Vor- und Nachteile haben. So kann die Lesbarkeit für Menschen erhöht werden, wenn in ein textbasiertes Format umgewandelt wird, dies hat jedoch oft eine Erhöhung der Dateigröße zur Folge. Zudem wird das Verhältnis von Inhalt zu Syntax schlechter, je umfangreicher Felddefinitionen sind. Daher benötigt beispielsweise XML üblicherweise mehr Speicherplatz als CSV.[9]

3 Datenformate

Um Informationen zu transportieren, ob in analoger oder digitaler Form, werden Dokumente erstellt. Diese können aus Text-, Bild-, Ton- oder Videodateien bestehen. Da diese Arbeit sich nur mit dem textuellen Datenaustausch beschäftigt, können andere Formate vernachlässigt werden. Grundsätzlich teilen sich Dokumente in Inhalt, Layout und die logische Struktur auf. Je nach Sinn und Zweck des Dokumentenaustauschs wurden und werden immer neue Formate entwickelt. Das Unternehmen .Versicherung verwendet vor allem die Datenformate GDV, XML, JSON, CSV und EDIFACT zur internen und externen Datenübertragung. Um ein grobes Verständnis für den Aufbau und die Funktion der einzelnen Formate zu vermitteln, werden diese im folgenden Kapitel genauer vorgestellt.

3.1 GDV-Satz

In der heutigen Zeit ist der „elektronische Datenaustausch“[10] ein wichtiger Faktor in Unternehmen. Auch in der Versicherungsbranche wird hier ein problemloser Ablauf angestrebt. Bei der internen Kommunikation ist es vergleichsweise einfach, da alle verwendeten Datenformate bereits in die verschiedenen Geschäftsprozesse integriert sind und somit optimal verwaltet werden. Sobald jedoch ein Austausch von Informationen zwischen verschiedenen Unternehmen stattfindet, kommt es oftmals zu Problemen auf Grund unbekannter oder nicht etablierter Datenformate. Daher findet in der Versicherungswirtschaft oftmals noch ein reger Papierverkehr statt, wodurch etwaige Fehler (Doppelerfassungen, Dateninkonsistenz) begünstig werden. Um dieses Problem zu beheben, hat der Gesamtverband der Deutschen Versicherungswirtschafts e.V. (GDV) eine Reihe von Normen entwickelt. Die erste Version

[9] Vgl. Hesse, F. (2015)
[10] Gesamtverband der Deutschen Versicherungswirtschaft e. V. (2013), S. 6

des sogenannten VU-Vermittlers wurde bereits in den 80er Jahren veröffentlicht. In dieser Satzung werden vor allem gewisse Standards festgehalten, die Datenstruktur, Aufbau, Größe und Syntax, sowie bestimmte Fachbegriffe und Geschäftsvorfälle beschreiben.

Inzwischen ist der GDV-Datensatz weit verbreitet und verbessert die Kommunikation zwischen Versicherungsunternehmen und Vermittlern in ca. 93% aller Betriebe in Deutschland. Mit Hilfe dieser Normen kann ein effizienter Datenaustausch von Geschäftsdaten, Bestandsdaten, Abrechnungsdaten und Schadensinformationen auch unternehmensübergreifend stattfinden. Das allgemeine Ziel, welches hier verfolgt und ebenfalls unter EDI erwähnt wird, ist eine Vollautomatisierung der Kommunikation. Bei der Erstellung des VU-Vermittlers wurden spezielle Grundsätze beachtet. Darunter fallen beispielsweise die Transformation von Papierverkehr zu elektronischen Nachrichtentypen, sodass nur ein einziges Konverterprogramm Verwendung findet, Soft- und Hardwareneutralität nach EDI, Flexibilität der Prozesse (Darstellungscodes, Medien, sowie Übermittlungswege), datenschutzrechtliche Aspekte, Investitionsschutz, etc.[11]

Grundsätzlich bezieht sich die vorgegebene Struktur des Datensatzes auf die Übertragung von „Bestands-, Inkasso-, und Schadensinformationsdaten [...][sowie den] Austausch von Geschäftsvorfalldaten wie eVB, Antrag und Vermittlerabrechnung“[12].

Der allgemeine Aufbau wird in Abb. 2Abb. 1 dargestellt.

Übertragungsdatei
Datenpaket (1-n)
Vorsatz – Satzart x001 (1-mal)
Datensätze (1-n)
Nachsatz – Satzart x999 (1-mal)

Abb. 1: Allgemeiner Aufbau - GDV

Die Übertragungsdatei stellt hier die technischen Spezifikationen des Datenaustauschs, aufgeteilt in einzelne Pakete, dar. Jedes einzelne Datenpaket wird wiederum durch den Vor- und Nachsatz genauer definiert. Diese werden durch spezielle Satzarten beschrieben und müssen grundsätzlich immer dann neu festgelegt werden, wenn sich bestimmte Informationen, wie beispielsweise die Vermittler-Nummer, ändern. Der Datensatz, welcher tatsächliche Informationen enthält, ist $256-256^n$ Bytes lang und wird ebenfalls durch Satzarten, wie z.B. 0300 für Beteiligungs-Informationen, bestimmt. Es existiert also ein gewisser Nummernkreis, welcher Vorsatz, Datensatz und Nachsatz festlegt, aufgeteilt in verschiedene themenspezifische Informationen, wie Partnerdaten, Allgemeiner Vertragsteil, Investmentfonds und viele

[11] Vgl. Gesamtverband der Deutschen Versicherungswirtschaft e. V. (2013), S.1 ff.
[12] Ebenda S. 14

mehr. Den einzelnen Datenfeldern werden hierbei gewisse Formate zugewiesen, die eingehalten werden müssen, d.h. z.B. das Datum soll numerisch angegeben werden im Format „TTMMJJJJ“. Hier unterscheidet man zwischen alphanumerischen und numerischen Eingaben. Die verwendete Syntax ist hier zu beachten: Die einzelnen Felder werden durch ihre Stelle im Dokument gekennzeichnet. Damit sich die Information am richtigen Ort befindet werden Leerzeichen (bei alphanumerischen Feldern) und Nullen (bei numerischen Feldern) als Platzhalter verwendet; es werden keine weiteren Trennungszeichen etc. benötigt. Neben dem Format werden die Felder auch in konstante Felder (Satznummer etc.), Nutz-Felder (Klartext, Betragsfelder, Schlüsselwerte), Key-Felder (Identifikationsfelder) und Referenzfelder (Verknüpfung einzelner Satzarten) eingeteilt.[13]

Wie ein Ausschnitt aus einem dieser Formulare im Endeffekt aussehen kann beschreibt folgendes Beispiel (Abb. 2):

```
00019999 XXX Versicherung AG          BRBRIENNEE,JÜRGEN           22072004220720049999009999
00019999 XXX Versicherung AG          BRBRIENNEE,JÜRGEN           22072004220720049999009999
01009999  030      599999999990199990099992Kitzelpfütze
01009999  030      59999999999019999009999          B4LTTT
02009999  030      59999999999019999009999201052004010520090105200511  0000000001        01052004
02109999  030      5999999999901999900999910105200401052009010520040901 000000000000000000 EUR00
02209999  030      5999999999901999900999900000011Kitzelpfütze                  000000Kitzelpfütze
02209999  030      5999999999901999900999900000120000000000113050000000000000141950000000000000
02219999  030      5999999999901999900999900000120000000500000000000025000000000000000000000000
02209999  030      599999999990199990099993                                   0000000000000000000
```

Abb. 2: Beispiel einer GDV Datei[14]

Hier sieht man deutlich die Satzart zum Beginn jeder Zeile, sowie die Trennung der Felder durch Leerzeichen und Nullen.

3.2 XML

Extensible Markup Language, kurz XML, ist ein Dateiformat, das 1996 von der XML Working Group entwickelt wurde. Die Entwickler haben versucht ein Format zu schaffen, welches benutzerfreundlich, einfach zu bedienen, und direkt über das Internet nutzbar ist. Neben diesen Design-Zielen gab es noch einige weitere, wie beispielsweise dass XML mit SGML kompatibel sein soll, dass es möglichst viele Anwendungen unterstützen soll, eine kurze und

[13] Vgl. Gesamtverband der Deutschen Versicherungswirtschaft e. V. (2013), S. 14 ff.
[14] Mit Änderungen entnommen aus: Gesamtverband der Deutschen Versicherungswirtschaft e. V. (2005)

klare Struktur haben und möglichst wenige bis gar keine optionalen Funktionen existieren sollen.[15]

Die erste Version von XML wird außerdem durch eine Reihe von Standardnormen was die Sprache und das Internet betrifft, definiert. So wird hier der ISO/IEC 10646 Standard für die Zeichen benutzt.[16]

All diese verschiedenen Merkmale definieren XML. Um ein Objekt als XML Dokument bezeichnen zu können muss es sowohl eine gewisse logische, als auch physische Struktur aufweisen. Allgemein wird dieser Zustand als „wohlgeformt“ bezeichnet. Hierbei ist es wichtig, dass das Dokument allen festgelegten Spezifikationen folgt, die von W3C festgelegt wurden.[17]

Die physische Struktur beschreibt hier den Aufbau aus verschiedenen Speichereinheiten. Ein XML Dokument besteht aus mindestens einer dieser Einheiten, die als „Entities“ bezeichnet werden, und jeweils einen eigenen Inhalt besitzen. Dieser ist entweder „parsed“, besteht also aus einem XML-Text und kann normal interpretiert werden, oder „unparsed“. Letzteres bedeutet, dass die Einheit z.B. ein Bild oder ein anderes Dokument ist, welches von XML nicht erkannt wird und auf Grund dessen über einen Parser/Prozessor verarbeitet werden müssen, mit Hilfe einer ihnen zugewiesenen Notation.[18] Zu Beginn jedes Dokuments steht die sogenannte „root“-Entität, welche als Startpunkt für das Verarbeitungsprogramm dient.[19]

Die logische Struktur hingegen beschreibt den konkreten Inhalt eines XML-Dokuments. Sie wird durch das Markup festgelegt und besteht aus Elementen, Kommentaren, Deklarationen und Charakter Referenzen, sowie Attributen.[20] Die einzelnen Elemente werden durch Marken (engl. „tags“) umschlossen. Auch die Struktur und Anordnung der einzelnen Elemente wird vom Markup festgelegt, wobei die genaue Ausgabe dieser Informationen vom ausgebenden Programm abhängig ist. Die allgemeine Syntax wird in jedem Dokument als DTD (Document Type Definition) im Prolog der Datei festgelegt. Hier wird beispielsweise die verwendete Version von XML festgelegt.[21]

Die Besonderheit von XML liegt in seiner Flexibilität. Diese wird durch die Entitäten erreicht, deren Inhalt sowohl innerhalb als auch außerhalb des Dokuments gespeichert sein kann.

[15] Vgl. Farsi, R. (1999), S. 436
[16] Vgl. Bray, T. u. a. (1998), S. 4
[17] Vgl. Bray, T. u. a. (1998), S. 4
[18] Vgl. Ebenda, S. 4
[19] Vgl. Ebenda, S. 34
[20] Vgl. Ebenda, S. 4
[21] Vgl. Farsi, R. (1999), S. 436

Der Inhalt bleibt also unverändert, wohingegen sich die äußere Form dank der Markups an das jeweilige Verarbeitungsprogramm anpasst.[22]

Folgende Abb. 3 stellt einen kurzen Prolog, sowie ein paar Zeilen XML dar:

```
< ?xml version = „1.0“ encoding =
„UTF-8“? >
< !DOCTYPE mail SYSTEM
„http://server. de/mail. dtd“ >
```

Abb. 3: XML Beispiel[23]

Im Prolog wird festgelegt, dass hier die Regelungen von XML Version 1.0 gelten und der Zeichensatz UTF-8 verwendet wird. Das Dokument ist in diesem Fall eine E-Mail, deren DTD extern auf einem Server zu finden ist.[24]

3.3 JSON

Die JavaScript Object Notation, JSON, wurde von der ECMAScript Programmiersprache abgeleitet. Es handelt sich also um ein relativ simples, textbasiertes und von anderen Programmiersprachen unabhängiges Datenaustauschformat. Es soll sowohl für Mensch als auch Maschine leicht zu lesen und zu schreiben sein. Zusammen mit XML ist es eines der am häufigsten verwendeten Datenaustauschformate für APIs. Der Trend zeigt, dass immer mehr APIs mit JSON kompatibel sind und dieses Format XML langsam verdrängt.[25]

JSON wird vor allem durch den einfachen Aufbau gekennzeichnet. Es gibt vier verschiedene Typen, Textstrings, Nummern, Boolean und Null, sowie zwei Strukturen: Objekte und Arrays. Im Folgenden wird kurz auf die einzelnen Typen eingegangen[26]:

- Textstring (String): Textstrings werden in JSON so ähnlich wie in Programmiersprachen die auf C basieren dargestellt. Dies bedeutet, dass jeder String mit Anführungszeichen beginnt und endet. Dazwischen kann jede beliebige Folge von Unicode-Zeichen stehen, ausgenommen Anführungszeichen, Backslash und die Unicode Kontrollzeichen. Diese Sonderfälle können jedoch so dargestellt werden das sie trotzdem ausgegeben werden.

[22] Vgl. Farsi, R. (1999), S. 436 f.
[23] Enthalten in: Farsi, R. (1999), S. 437
[24] Vgl. Farsi, R. (1999), S. 437
[25] Vgl. Siriwardena, P. (2014), S. 201
[26] Vgl. Bray, T. (2014), S. 5 ff.

- Nummer (Number): Die numerische Darstellung in JSON gleicht stark der in anderen Programmiersprachen. So befindet man sich in einem Dezimalsystem, d.h. die Basis ist 10. Durch plus oder minus kann die Zahl genauer beschrieben werden. Es ist ebenfalls möglich Teilwerte oder Exponenten darzustellen. Insgesamt gleicht die Schreibweise der, welche auch in JavaScript verwendet wird. Bei JSON ist es wichtig auf die Genauigkeit zu achten, da diese nicht von allen Systemen unterstützt wird. Ebenso ist es nicht möglich numerische Werte mit einer „0" einzuleiten.
- Boolean: Wie bereits aus anderen Standards bekannt, kann ein Boolean exakt zwei Zustände annehmen: true oder false, also richtig oder falsch bzw. ja oder nein.
- Null: Null ist ein nicht änderbarer Zustand. Er bezeichnet nichts bzw. etwas Leeres.
- Objekt (Object): Objekte werden durch geschweifte Klammern dargestellt, die entweder keinen, einen oder mehrere Namen bzw. Werte enthalten. Hierbei ist es wichtig, dass die Namen in einem Objekt einzigartig sind, da ansonsten die Interoperabilität gefährdet ist. Innerhalb des Objektes werden Namen durch Kommata von Werten und Werte durch einen Doppelpunkt von Namen getrennt.
- Array: Bei einem Array werden kein, ein oder mehrere Elemente bzw. Werte von eckigen Klammern umschlossen. Bei einer Auflistung mehrere Elemente werden diese durch Kommata getrennt. Die Werte oder Elemente müssen nicht dem gleichen Typus angehören. Es sind somit keine weiteren Regelungen in JSON festgelegt.

JSON verwendet standardmäßig eine UTF-8 Kodierung, sowie Unicode, um eine hohe Interoperabilität zu gewährleisten. Ausnahmen sind hier eher selten und können zu unvorhersehbaren Reaktionen der verarbeitenden Software führen. Da JSON wie bereits erwähnt auf ECMA-Script basiert und größtenteils die Syntax von JavaScript verwendet, wird es oft zum Datenaustausch zwischen Applikationen die in folgenden Programmiersprachen geschrieben sind benutzt: ActionScript, C, C#, Clojure, ColdFusion, Common Lisp, E, Erlang, Go, Java, JavaScript, Lua, Objective CAML, Perl, PHP, Python, Rebol, Ruby, Scala, und Scheme[27].

Durch die einfache Darstellung und Handhabung von JSON wurde das ursprüngliche Ziel einer minimalistischen, leicht übertragbaren und textbasierten Programmiersprache erfolgreich umgesetzt.[28]

Anhand eines Beispiels auf der nächsten Seite (Abb. 4) werden die einzelnen Komponenten von JSON nochmal aufgegriffen und dargestellt:

[27] Vgl. Bray, T. (2014), S. 9 ff.
[28] Vgl. Bray, T. (2014), S. 1-16

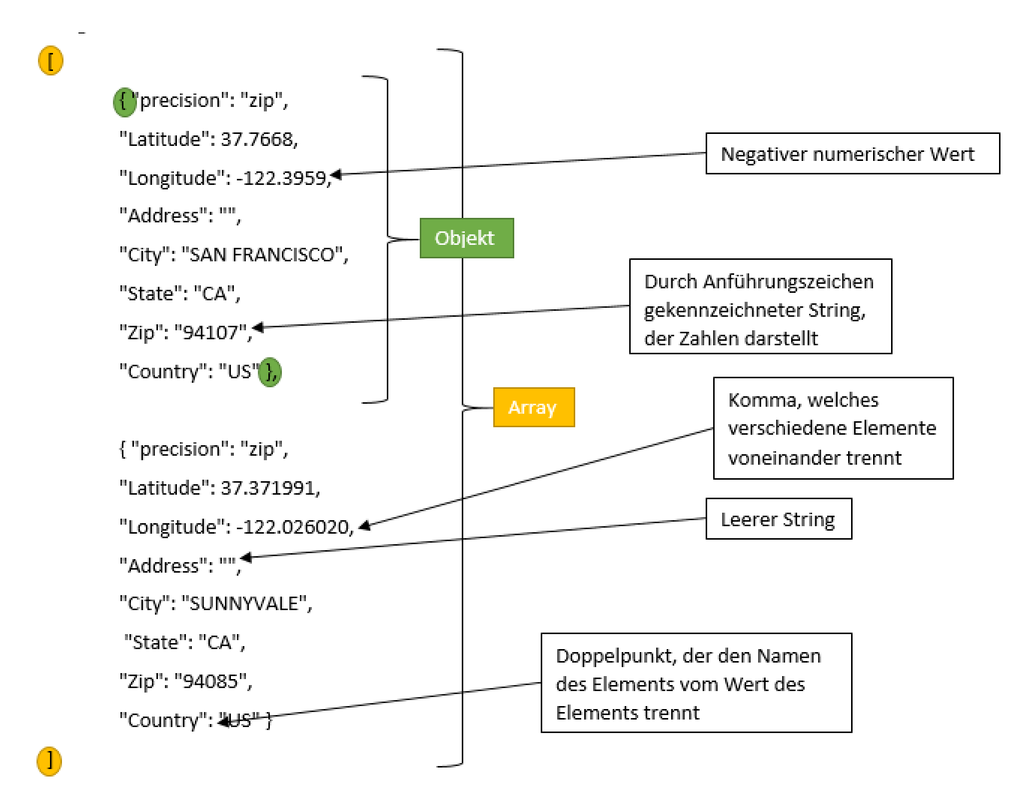


Abb. 4: JSON Beispiel[29]

3.4 CSV

Das Textformat CSV (comma separated values) ist eine sehr simple Darstellung von Informationen. Das Format wurde entwickelt um den Austausch und die Konvertierung von Daten zwischen verschiedenen Tabellenkalkulationssystemen so einfach und sicher wie möglich zu gestalten. Hierfür eignet sich CSV immer noch am besten. Inzwischen wird es allerdings auch für den Austausch simpler Textdateien verwendet.[30]

Obwohl das Format sehr weit verbreitet ist, existiert kein offizieller Standard. Es wurde bis jetzt nur durch IETF genauer beschrieben und definiert. Dies führt häufig zu Missverständnissen und unterschiedlichen Darstellungen bei der Interpretation der Daten oder der Konvertierung. Im Folgenden soll das Format so beschrieben werden, wie es heutzutage zum Großteil verwendet wird.[31]

Der am häufigsten verwendete Zeichensatz ist die ISO 8859 Kodierung, in seltenen Fällen auch Unicode. Die Syntax bei CSV ist sehr einfach strukturiert: CSV-Dateien bestehen aus einem oder mehreren Einträgen, die durch Zeilenumbrüche getrennt werden. Es besteht die Möglichkeit, durch eine erklärende Zeile zu Beginn des Dokuments die einzelnen Felder in

[29] Mit Änderungen entnommen aus: Bray, T. (2014), S. 12
[30] Vgl. Hoffmann-Walbeck, T. u. a. (2013), S. 33 f.
[31] Vgl. Shafranovich, Y. (2005): S. 2

den Zeilen genauer zu beschreiben. Dies ist jedoch nicht erforderlich. Jede Zeile beinhaltet beliebig viele Felder, die durch Kommata getrennt werden. Idealerweise beinhaltet jede Zeile in einem Dokument die gleiche Anzahl an Feldern. Um sicherzustellen, dass die Werte innerhalb der Felder korrekt interpretiert werden, können sie von Anführungszeichen umschlossen werden. So lassen sich beispielsweise ohne Missverständnisse Kommata oder Zeilenumbrüche als Feldinhalt darstellen. Nach der letzten Zeile ist kein Zeilenumbruch nötig. Außerdem zu beachten ist, dass am Ende einer Zeile kein Komma nach dem letzten Feldeintrag stehen darf. Leerzeichen gehören zum Textinhalt und jegliche Gestaltung der Daten (Farbe, bold, italic, usw.) geht bei der Transformation in oder aus dem CSV Format verloren.[32]

Im folgenden Beispiel wird ein einfacher Datensatz im CSV-Format dargestellt:

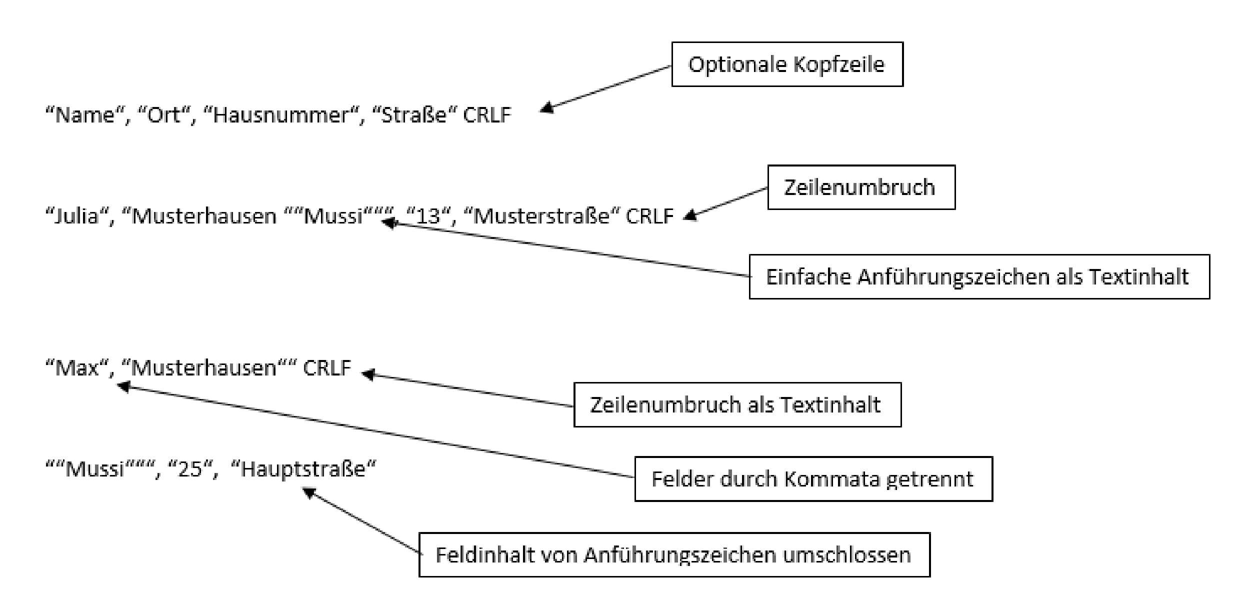


Abb. 5: CSV Beispiel

Ausgabe der Datei als Spreadsheet in Excel (bei der Umwandlung von *.csv in *.xlsx sind keine doppelten Anführungszeichen nötig, insofern sie nicht einen Zeilenumbruch innerhalb eines Feldes signalisieren oder ein Anführungszeichen ausgegeben werden soll):

Eingabetext in *.csv:

Name, Ort, Hausnummer, Straße

Julia, "Musterhausen ""Mussi""",13, Musterstraße

Max, "Musterhausen""

Mussi""", 25, Hauptstraße

Ausgabetabelle in *.xlsx:

	A	B	C	D
1	Name	Ort	Hausnummer	Straße
2	Julia	Musterhausen "Mussi"	13	Musterstraße
3	Max	Musterhausen" Mussi"	25	Hauptstraße

[32] Vgl. Hoffmann-Walbeck, T. u. a. (2013), S. 33 f.

3.5 EDIFACT

Bevor der Datentyp EDIFACT genauer vorgestellt wird, ist es wichtig eine Wissensgrundlage zum Thema EDI – Electronic Data Interchange – zu schaffen.

Der Kerngedanke von EDI ist es, jeglichen Papierverkehr innerhalb eines Unternehmens zu reduzieren bzw. gänzlich abzuschaffen und durch elektronischen Datenaustausch zu ersetzen. Hier ist EDI mit gewöhnlichen E-Mails zu vergleichen. Der größte Unterschied ist, dass E-Mails einen Datenaustausch von Mensch zu Mensch über eine Maschine bewerkstelligen, wohingegen EDI einen direkten Datenaustausch zwischen zwei Maschinen beschreibt. Um dies möglich zu machen, müssen sowohl Sender als auch Empfänger den gleichen Standards folgen und die Daten in einer strengen, vorgegebenen Struktur zur Verfügung stellen.[33]

Das erste Dokument zum standardisierten Datenaustausch wurde 1975 von TDCC, dem Transportation Data Coordinating Committee, unter dem Namen EDI veröffentlicht und beschrieb hauptsächlich eine Struktur zur Übertragung von Luft-, Motor-, See- und Zugtransportdokumenten. Diese anfängliche Ansammlung von Standards wurde im Laufe der Zeit erweitert und schließlich unter UN/EDIFACT (United Nations Directories for Electronic Data Interchange for Administration, Commerce and Transport) als weltweit gültige Norm zusammengefasst.[34] Neben EDIFACT als Normensammlung wird bei EDI-Standards beispielsweise auch X12 oder VDA verwendet.[35] Und obwohl EDIFACT nach dem Aufkommen von XML einen deutlichen Nutzungsrückgang zu verzeichnen hatte, so ist es dennoch aus der Verwaltungs-, Handels- und Transportbranche nicht mehr wegzudenken. Dafür ist vor allen Dingen die frühe ISO Normierung verantwortlich und die viele Arbeit, die es bedeuten würde einen neuen Standard auf den Informations- und Anpassungslevel von EDIFACT zu bringen.[36]

Im weiteren Verlauf wird nun näher auf die Datenstruktur von EDIFACT eingegangen. Ähnlich wie beim GDV-Datenformat werden auch hier für verschiedenste Informationen Codes verwendet, wie beispielsweise Ländercodes oder Codes zur Bestimmung des Datumformats. Jedoch werden einzelne Felder in EDIFACT nicht durch ihre Position, sondern durch gewisse Trennzeichen voneinander abgegrenzt. Insgesamt kann man das Format in vier Teile unterteilen: Syntax, Datenelemente (kleinste Einheiten), Segmente (bestehen aus einer

[33] Vgl. Unitt, M./ Jones, I. (1999), S. 17
[34] Vgl. Balsmeier, P./ Borne, B. (1995), S. 53 f.
[35] Vgl. GEFEG mbH (2014)
[36] Vgl. Unitt, M./ Jones, I. (1999), S. 4

Gruppe gleichartiger Datenelementen) und Nachrichten (geordnete Sequenz von Segmenten um beispielsweise einen Lieferschein darzustellen).[37]

Der grobe Aufbau einer Datei im EDIFACT-Format wird am folgenden Bestellscheinbeispiel genauer erklärt:

EDIFACT
Muster einer Bestellung per EDIFACT Seite 1 edifact

Kopf	**UNB+UNOA:1+MAURIKS +AMF+020131:1700+AMF00001'** **UNH+00001+ORDERS:D:93A:UN:EAN007'** **BGM+220+5761650'** **DTM+4:20020131:102'** **NAD+BY+30809'** **NAD+SU+02/0519'**	
Positionen	**LIN+1'** **PIA+1+140327:SA+27408103:BP'** **QTY+21:100:PCE'** **DTM+2:20020131:102'**	*1. Pos.*
	LIN... **PIA...** **OTY...** **DTM...**	*2. Pos. usw.*
Ende	**UNS+S'** **UNT+15+00001'** **UNZ+1+AMF00001'** ↑ *Segmente*	

Abb. 6: EDIFACT Beispiel[38]

Auf den ersten Blick zu erkennen ist die Verwendung von Satzzeichen. In GDV-Dateien werden einzelne Segmente durch ein einfaches Hochkomma beendet (hier am Ende jeder Zeile zu finden). Das Pluszeichen separiert einzelne Datensegmente voneinander, welche wiederum durch Doppelpunkte in kleinere Einheiten unterteilt werden können.[39]

Im Kopfteil des Dokuments werden die wichtigsten Daten festgelegt. UNB beschreibt hier die Übertragungsdatei genauer im Hinblick auf Absender, Empfänger, Datum der Erstellung, etc. Hierbei werden sowohl numerische Zeichenfolgen, als auch alphanumerische verwendet. Im UNH-Segment wird neben der Versionsnummer auch der Typ der Nachricht festgelegt, welcher in diesem Fall eine Bestellung („ORDER") ist. Es existieren viele solcher Segmente, welche teilweise zwingendermaßen im Dokument deklariert werden müssen, so wie UNB

[37] Vgl. Ecosio GmbH (2015)
[38] Enthalten in: Häge, M. (o. J.), S. 1
[39] Vgl. UN/CEFACT Syntax Working Group (1998), S. 3

und UNH. Des Weiteren wird im Kopf der Datei auf das Datum und spezifische Informationen zur Nachricht (Identifikationsnummer, Name, Bestellnummer, Adressen) eingegangen. Der Mittelteil beschäftigt sich hingegen hauptsächlich mit der Darstellung und Beschreibung der verschiedenen Produkte. Zum Schluss ist es wichtig festzuhalten, wie viele Segmente das Dokument aufweist. Das Segment UNZ dient der Beendigung der Übertragung nachdem sie auf Vollständigkeit geprüft wurde.[40]

Im Umfang dieser Arbeit und im Hinblick auf das Ziel derselben ist es nicht notwendig, alle Kürzel und syntaktischen Feinheiten des GDV-Formats zu erläutern. Es ist wichtig einen Überblick über die Darstellung und grobe Funktion zu erhalten, um im folgenden Verlauf der Arbeit die Verwendung des Begriffes GDV nachzuvollziehen.

4 Kriterienkatalog

Der zentrale Punkt bei der Auswahl möglicher Produkte ist der Kriterienkatalog. Mit diesem soll sichergestellt werden, dass alle relevanten Kriterien berücksichtigt worden sind. Basierend auf dem Kriterienkatalog und der späteren Gewichtung der Kriterien im Rahmen der Auswahlverfahren werden die Produkte analysiert und bewertet. Dabei wird sowohl die Auswahl der Kriterien, als auch deren späteren Gewichtung an die Anforderungen des Unternehmens individuell angepasst, um ein möglichst optimales Ergebnis zu erhalten.

Wichtig bei diesem Verfahren ist, dass die Kriterien möglichst einfach und nachvollziehbar sind. Deshalb liegt der Schwerpunkt des Kriterienkatalogs nicht auf der Vollständigkeit der Kriterien, sondern bezieht sich in erster Linie auf die geforderten Schwerpunkte. Das bedeutet, dass er mit seinem pragmatischen Aufbau nicht das beste Produkt finden soll, das es auf dem Markt gibt, sondern jenes, welches den geplanten Einsatzzweck am besten erfüllt. Dementsprechend müssen als erstes die Grundlagen und Spezifikationen für die Auswahl der Kriterien festgelegt werden.

[40] Vgl. UN/CEFACT Syntax Working Group (1998), S. 19 ff.

4.1 Auswahl der Kriterien

Die Auswahl der Kriterien basiert auf Recherche praxisnaher Literatur sowie auf einem Gespräch mit den Verantwortlichen bei der .Versicherung, welche auch die Zielgruppe dieser Arbeit ist.

Die so gewonnen Erkenntnisse werden nachfolgend aufbereitet und in den entsprechenden Bereichen erklärt. Bei der Auswahl der Kriterien wurde darauf geachtet, dass die einzelnen Merkmale sich nicht überschneiden und so unabhängig voneinander bewertet werden können. Grund dafür ist, dass eine Überschneidung der Kriterien zu einer Verfälschung des Endergebnisses führen kann.

4.2 Aufbau des Kriterienkatalogs

Für ein besseres Verständnis sind die Kriterien in vier Kategorien eingeteilt. Teil A befasst sich mit den Kosten und allgemeinen funktionalen und systemtechnischen Rahmenbedingungen. Im Wesentlichen wird hier beschrieben, welche Kosten pro Lizenz zu erwarten sind und welche Betriebssysteme von den Transformationstools unterstützt werden. Teil B befasst sich mit der funktionalen Abdeckung. Hier wird sowohl überprüft, ob das Produkt mit den geforderten Datentypen arbeiten kann, als auch, ob es weitergehende Spezifikationen in Form eines Regelwerks zulässt. Teil C umfasst branchenspezifische Funktionalitäten, wie zum Beispiel die Einbindung in andere Anwendungen und Prozesse oder das Logging und Monitoring.
Im letzten Teil, Teil D, sind anbieterbezogene Kriterien beschrieben, sprich der Service. Das hat vor allem den Grund, das Produkt dahingehend zu untersuchen, ob es noch weiterentwickelt wird und ob es entsprechenden Support dafür gibt.

Zur Verdeutlichung der Kategorien zeigt die nachfolgende Graphik (Abb. 7 nächste Seite) die Verteilung der Kriterien. (X)

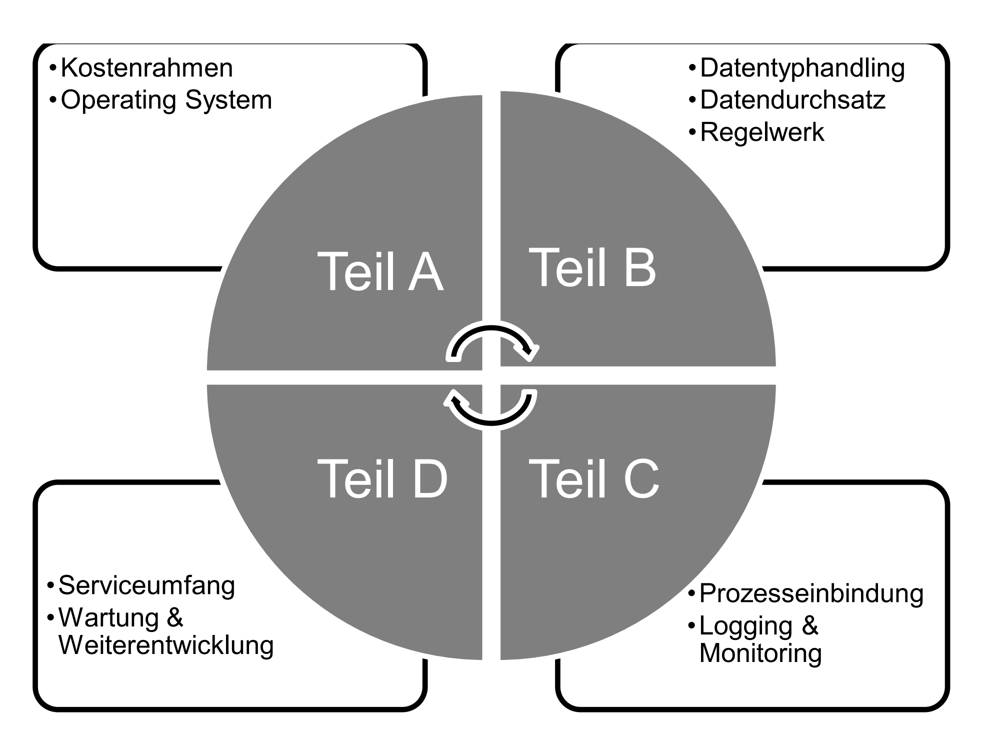


Abb. 7: Zusammensetzung Kriterienkatalog

Resultierend aus der vorigen Auswahl ergibt sich nachfolgender gültiger Kriterienkatalog. Die Tab. 1 auf der nächsten Seite gibt einen Überblick über die ausgewählten Kriterien und deren Beschreibung.

Kriteriengruppe	Kriterium	Beschreibung
Rahmenbedingungen	Kostenrahmen	Betrachtet werden hier ausschließlich die Anschaffungskosten der Software, Kosten für Service oder Support werden nicht mit eingerechnet.
	Operating System	Hier wird untersucht welche Betriebssysteme von der Software unterstützt werden. Es werden sowohl serverseitige, als auch clientseitige Betriebssysteme betrachtet.
Funktionale Abdeckung	Datentyphandling	Ein wichtiges Kriterium ist die Fähigkeit der Tools eine große Anzahl an Datentypen verarbeiten zu können. Relevant für die Hallesche sind vor allem die Datentypen XML/XSD, JSON, CSV, EDIFACT, und der GDV Satz.
	Datendurchsatz	Das Kriterium Datendurchsatz umfasst mehrere Dimensionen. Zum einen wird die Fähigkeit betrachtet eine hohe Datenmenge verarbeiten zu können, zum anderen spielt die Verarbeitungszeit eine wichtige Rolle. Es wird angenommen, dass der Datendurchsatz bei Tools mit „Batchmode“ höher ist, als bei Tools, welche auf manuelles Einlesen der Daten angewiesen sind.
	Regelwerk	Es wird betrachtet, ob Regeln für die Datentransformation aufgestellt werden können, und ob diese via Code oder GUI eingegeben werden.
Branchenlösungen	Prozesseinbindung	Prozesseinbindung beschreibt die Fähigkeit die Software in andere Applikationen oder Prozesse einbinden zu können, beispielsweise in den jeweiligen E-Mail-Client.
	Logging & Monitoring	Hier wird betrachtet, wie Fehler gehandelt, Transaktionen gemonitored und gelogged werden.
Service	Serviceumfang	Wie umfangreich ist der Service? Bietet der Hersteller des Tools beispielsweise Support bei der Einrichtung, bei Problemen, oder bei der Wartung der Software an?
	Wartung & Weiterentwicklung	Es wird untersucht, ob für das Programm eine aktive Entwicklergemeinschaft existiert. Dadurch soll verhindert werden, dass eine Software eingesetzt wird für welche keine Updates mehr erscheinen.

Tab. 1: Kriterienkatalog

5 Bewertung der Open Source - Transformationstools

Um aus der Vorauswahl aus mehreren OS-Transformationstools die drei besten Softwarelösungen herauszufinden, wird die Analyse und Bewertung der Produkte auf zwei Schritte aufgeteilt. Zunächst werden im ersten Schritt alle Produkte anhand einer Nutzwertanalyse verglichen und nach ihrer erreichten Punktzahl sortiert. Anschließend werden die drei am höchsten bewerteten Produkte im zweiten Schritt mit Hilfe der Analytic Hierarchy Process-Analyse (kurz: AHP-Analyse) erneut miteinander verglichen, um die finale Reihenfolge zu ermitteln. Im Folgenden werden die beiden Analysemethoden, die Nutzwertanalyse und die AHP-Analyse, erklärt und deren Anwendung und Ergebnisse beschrieben.

5.1 Nutzwertanalyse

5.1.1 Beschreibung

Die Nutzwertanalyse ist im Europäischen Raum, vor allem aber in der DACH-Region, eine anerkannte und häufig verwendete Methode um Entscheidungsfindungen zu unterstützen. Das Ziel der Nutzwertanalyse ist eine objektive Messung von Entscheidungsmöglichkeiten auf Basis einer quantitativen, nicht monetären Bewertung. Dabei können sowohl qualitative, als auch quantitative Kriterien betrachtet werden. Die Nutzwertanalyse findet vor allem dann Anwendung, wenn die Anzahl der entscheidenden Kriterien zu hoch und zu komplex ist, um eine Entscheidung durch logischen Verstand herbeiführen zu können. [41]
Ein wesentlicher Vorteil der Nutzwertanalyse ist zum einen ihre leichte Verständlichkeit und zum anderen ihre Flexibilität hinsichtlich der Anzahl der Kriterien und der Zielsetzung der Analyse. Jedoch ist die Bewertung stark von der objektiven Bewertung der Teilnehmer abhängig, weshalb eine gewisse Manipulationsgefahr und folglich ein erhöhtes Risiko der Inkonsistenz des Evaluationsprozesses entsteht. Zudem berücksichtigt die Nutzwertanalyse zwar quantitative und qualitative Merkmale, allerdings müssen quantitative Kriterien zunächst in eine „quasi-metrische“ Form gebracht werden. Dies kann zu Informationsverlust führen.[42]

5.1.2 Entscheidung

Nach Meinung der Autoren überwiegen die Vorteile bei achtsamer und korrekter Durchführung die Nachteile der Nutzwertanalyse. Vor allem in Hinblick auf die hohe Anzahl von alternativen Entscheidungsmöglichkeiten (insgesamt sieben) bringt die Simplizität der Nutz-

[41] Vgl. Kühnapfel, J. (2014), S. 1 ff.
[42] Vgl. Riedl, R. (o. J.), S. 115 ff.

wertanalyse einen entscheidenden Vorteil. Daher halten wir die Nutzwertanalyse für eine geeignete Methode, um die Vorauswahl von sieben verschiedenen OS-Transformationstools zu bewerten.

5.1.3 Durchführung

Die Durchführung der Nutzwertanalyse ist im Fall der hier genannten Entscheidungsfindung in fünf grobe Schritte aufzuteilen. Diese Schritte werden im Folgenden genauer erläutert.[43]

1. Benennung des Entscheidungsproblems
2. Auswahl der Entscheidungsalternativen
3. Sammlung der Entscheidungskriterien
4. Gewichtung der Entscheidungskriterien
5. Bewertung der Entscheidungsalternativen & Berechnung des Nutzwertes

Benennung des Entscheidungsproblems

Generell können für die Nutzwertanalyse zwei Arten von Entscheidungsproblemen identifiziert werden – das Treffen einer Entscheidung bei Auswahlproblemen und die Priorisierung von Auswahlmöglichkeiten.[44]
Wie schon in Kapitel 1 beschrieben verfolgt diese Arbeit den Zweck, ein geeignetes Open Source Datentransformationstool für die .Versicherung zu finden. Da in diesem Rahmen eine hohe Anzahl an OS-Transformationstools zum Vergleich steht, ist das Ziel der hier beschriebenen Nutzwertanalyse eine Priorisierung, bzw. eine Rangliste der Auswahlmöglichkeiten zu erstellen.

Auswahl der Entscheidungsalternativen

Die zum Vergleich stehenden Entscheidungsalternativen wurden auf Basis einer umfangreichen Internetrecherche herausgesucht. Für die Vorauswahl wurden explizit proprietäre Datentransformationstools, wie z.B. *Stylus Studio* oder *XMLSpy*, und sogenannte ETL-Tools

[43] Vgl. Kühnapfel, J. (2014), S. 6
[44] Vgl. ebenda, S. 6 f.

(Abkürzung für *„Extract-Transform-Load*“) ausgegrenzt.
Es konnten schließlich für die Nutzwertanalyse folgende Open Source Datentransformations-tools identifiziert werden:

- JAXMLP
- Talend
- Bots
- OpenRefine
- Edival
- XmlGrid.net
- Web Karma

<u>Auswahl der Entscheidungskriterien</u>

Die Auswahl der Entscheidungskriterien erfolgte, wie bereits in Kapitel 4) erwähnt, basierend auf einem Gespräch mit Frau Gross von der .Versicherung und der Projektbeschreibung. Der Kriterienkatalog für die Nutzwertanalyse umfasst insgesamt neun Kriterien, welche zur besseren Veranschaulichung noch einmal kurz aufgelistet werden.

Kriteriengruppe	**Kriterien**
Rahmenbedingungen	Kostenrahmen
	Operating System
Funktionale Abdeckung	Datentyphandling
	Datendurchsatz
	Regelwerk
Branchenlösungen	Prozesseinbindung
	Logging & Monitoring
Service	Serviceumfang
	Wartung & Weiterentwicklung

<u>Gewichtung der Entscheidungskriterien</u>

Da die Kriterien angesichts des Entscheidungsproblems selten die gleiche Bedeutung besitzen, werden diese in der Regel unterschiedlich gewichtet.
Um eine möglichst objektive Gewichtung zu erzielen, bieten sich generell zwei Verfahren an

– die Gewichtung mit Hilfe von Kriteriengruppen und die Paarvergleichsmethode. Bei der ersten Variante werden die Kriterien zunächst in Gruppen sortiert, innerhalb der Gruppen bewertet, und anschließend mit der Gewichtung der jeweiligen Gruppe verrechnet. Im Vergleich dazu wird bei der Paarvergleichsmethode jedes Kriterium miteinander verglichen und gewichtet. [45]

Da nach unserer Meinung die Gewichtung anhand der Paarvergleichsmethode transparenter erscheint, wurde diese Variante gewählt. Von jedem Teilnehmer der Analyse wird die individuelle Gewichtung eines Kriteriums bewertet und in die Kreuztabelle eingetragen (siehe Abb. 8). Anschließend wird aus der Summe der Ergebnisse von jeder einzelnen Spalte die prozentuale Gewichtung der Kriterien errechnet.

Am Beispiel des Kriteriums „Kostenrahmen“ lässt sich dabei erkennen:

- Es ist gleich wichtig wie „Operating System“ und „Serviceumfang“.
- Es ist unwichtiger als „Datentyphandling“, „Datendurchsatz“, „Regelwerk für Transformation“, „Logging und Monitoring“, und „Wartung und Weiterentwicklung“.
- Es ist wichtiger als „Prozesseinbindung“

	Kriterium	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	*Summe*	*Gewichtung*
(I)	Kostenrahmen		2	0	1	0	3	0	2	0	*8*	*5.56*
(II)	Operating System	2		0	0	0	2	1	2	1	*8*	*5.56*
(III)	Datentyphandling	4	4		2	2	4	4	3	3	*26*	*18.06*
(IV)	Datendurchsatz	3	4	2		2	3	4	4	3	*25*	*17.36*
(V)	Regelwerk für Transformation	4	4	2	2		3	4	4	2	*25*	*17.36*
(VI)	Einbindung in andere Prozesse	1	2	0	1	1		2	2	1	*10*	*6.94*
(VII)	Logging- und Monitoring Komponente	4	3	0	0	0	2		3	1	*13*	*9.03*
(VIII)	Serviceumfang	2	2	1	0	0	2	1		0	*8*	*5.56*
(IX)	Wartung und Weiterentwicklung	4	3	1	1	2	3	3	4		*21*	*14.58*
											144	*100*

Abb. 8: Paarvergleich der Kriterien

<u>Bewertung der Entscheidungsalternativen & Berechnung des Nutzwertes</u>

Im nächsten Schritt wird die Zielerfüllung der sieben Entscheidungsalternativen ermittelt. Hierfür hat sich als Skala ein Bewertungskorridor von eins bis sechs als geeignet herausgestellt. Generell entspricht diese Skala der Schulnotenskala („1“ = „sehr gut“/ „6“ = „ungenügend“), jedoch ergibt sich für jedes Kriterium ein unterschiedlicher Bewertungsmaßstab:

[45] Vgl. Kühnapfel, J. (2014), S. 10 ff.

Kostenrahmen	1= kostenlos 2-3= 1-100€ 4-5= 101-1000€ 6= mehr als 1000€
Operating System (kurz: OS)	1= mehr als 3 verschiedene Server- & Client OS 2-3= mind. 3 verschiedene OS 4-5= mind. 2 verschiedene OS 6= reine Webapplikation
Datentyphandling	1-2= viele Datentypen, komplexe Transformation 3-4= viele Datentypen, einseitige Transformation 5-6= wenig Datentypen
Datendurchsatz	1-2= hoher Datendurchsatz 3-4= mittlerer Datendurchsatz 5-6= niedriger Datendurchsatz
Regelwerk für Transformation	1-2= Regeln durch GUI festgelegt 3-5= Regeln durch Code festgelegt 6= kein Regelwerk
Prozesseinbindung	1= möglich 6= nicht möglich
Logging & Monitoring	1-2= automatisches Logging, Monitoring, & Debugging 3-5= Fehlerausgabe 6= kein Fehlerhandling, kein Logging & Monitoring
Serviceumfang	1-2= viele zusätzliche Services 3-5= zusätzliche Services 6= keine Services
Wartung & Weiterentwicklung	1-2= aktive Entwicklergemeinschaft, regelmäßigen Updates 3-4= aktive Entwicklergemeinschaft, unregelmäßigen Updates 6= inaktive Entwicklergemeinschaft, keine Updates

Tab. 2: Bewertungsmaßstab für die Kriterien

Um den Nutzwert zu bestimmen, wird zunächst für jede Entscheidungsalternative die Bewertung der Merkmalsausprägung ermittelt. Tab. 3 veranschaulicht die Merkmalsausprägung jedes einzelnen OS-Transformationstools. Sollte die Bewertung eines Kriteriums auf Grund mangelnder Informationen nicht möglich sein, wird dieses Kriterium mit dem Wert „0“ bewertet.

Im nächsten Schritt wird nun die Gewichtung der Kriterien mit der Merkmalsausprägung jeder Entscheidungsalternative multipliziert. Es ist zu beachten, dass aufgrund der geringen Skalenspreizung der Schulnotenskala, d.h. der geringen Differenz aus höchster und niedrigster Bewertungseinheit, die Bewertung mit dem umgekehrten Schulnotenwert multipliziert werden sollte. Im Anschluss werden diese Ergebnisse für jede Entscheidungsalternative aufsummiert und bilden den Nutzwert.[46]

5.1.4 Ergebnis der Nutzwertanalyse

Wie in Tab. 3 (siehe nächste Seite) abzulesen ist, kann der maximal zu erreichende Nutzwert nicht mehr als 600 Punkte übersteigen. Nach Bewertung der sieben Open Source Datentransformationstools lässt sich feststellen, dass vier der analysierten Tools mehr als $\frac{2}{3}$ der maximalen Punktzahl, sprich mehr als 400 Punkte, erreicht haben. Diese vier Tools zeigen im Durchschnitt in allen Kategorien eine gute bis sehr gute Merkmalsausprägung. Vor allem die Programme *Talend*, *Bots* und *Web Karma* heben sich stark von der Konkurrenz ab. Sie stechen besonders dadurch hervor, dass sie eine Vielzahl von Datentypen unterstützen, einen hohen Datendurchsatz versprechen, oder dank einer aktiven Entwicklergemeinschaft regelmäßig gewartet und upgedatet werden. Das Schlusslicht der analysierten Datentransformationstools bilden die Tools *Edival, JAXMLP,* und *XmlGrid.net*. Diese Tools fallen auf Grund geringer Funktionalitäten und einer inaktiven Entwicklergemeinschaft aus dem Fokus.

[46] Vgl. Kühnapfel, J. (2014), S. 16 f.

Kriterium	Gewichtung in %	Alternativen						
		Talend	Bots	Web Karma	OpenRefine	XmlGrid.net	JAXMLP	Edival
Kostenrahmen	5.56	1	1	1	1	1	1	1
Operating System	5.56	1	1	2	2	6	2	0
Datentyphandling	18.06	1	1	2	2	4	6	0
Datendurchsatz	17.36	1	3	2	4	6	0	0
Regelwerk	17.36	1	3	2	3	6	0	0
Prozesseinbindung	6.94	1	1	1	6	6	0	0
Logging und Monitoring	9.03	1	2	3	3	5	0	0
Serviceumfang	5.56	1	1	3	1	6	6	6
Wartung und Weiterentwicklung	14.58	1	1	1	1	5	6	6
Summe (insg. 600 Punkte)		600	522	513	437	188	99	53

Tab. 3: Nutzwertanalyse OS-Transformationstools

5.2 Analytic Hierarchy Process

5.2.1 Beschreibung

Der Analytic Hierarchy Process, oder auch Analytische Hierarchieprozess genannt, wurde vom Mathematiker Thomas L. Saaty entwickelt und wird ähnlich wie die Nutzwertanalyse bei der Findung von komplexen Entscheidungsproblemen eingesetzt.
Bei der AHP-Methode wird das Entscheidungsproblem in mehrere Teilprobleme dekomponiert und in einer hierarchisch strukturierten Form dargestellt, wodurch die Komplexität der Entscheidungsfindung reduziert und die Transparenz erhöht wird. Für die Auswahl der optimalen Lösung des Entscheidungsproblems werden sowohl die Entscheidungskriterien, als auch die Entscheidungsalternativen miteinander verglichen und mit Hilfe von Matrizenalgebra berechnet. [47]
Ein Vorteil des AHP ist die Möglichkeit, quantitative und qualitative Bewertungsmerkmale direkt zu vergleichen. Verglichen mit der Nutzwertanalyse wird beim AHP-Vorgehen in diesem Punkt sichergestellt, dass es zu keinem Informationsverlust kommt. Ein weiterer Vorteil des AHP ist zudem die Überprüfung des Entscheidungsprozesses auf seine Konsistenz, d.h.

[47] Vgl. Banai-Kashani, R. (o. J.), S. 685 ff.

die Bewertung/ Gewichtung der Kriterien und Entscheidungsalternativen wird auf Widersprüche geprüft und dadurch die Validität des Entscheidungsprozesses garantiert.
Dies führt allerdings auch dazu, dass der AHP deutlich komplexer und weniger leicht verständlich ist als die Nutzwertanalyse.[48]

5.2.2 Entscheidung

Der Analytic Hierarchy Process bietet unserer Meinung nach einen guten Rahmen, um die vorhergegangene Nutzwertanalyse zu ergänzen und die Ergebnisse zu verifizieren. Mit Hilfe der AHP-Analyse sollen die drei OS-Transformationstools *Talend*, *Bots*, und *Web Karma*, welche laut der Nutzwertanalyse die Ränge eins bis drei belegen, erneut verglichen werden und in eine finale Reihenfolge gebracht werden.
Auf Grund des geringen Informationsverlusts und der hohen Konsistenz des Entscheidungsprozesses beim AHP-Vorgehen erwarten wir eine präzise Einschätzung der Entscheidungsalternativen.

5.2.3 Durchführung

Im Wesentlichen umfasst der Analytic Hierarchy Process drei Schritte – 1) Definition der Problemlösung und der Zielhierarchie, 2) Gewichtung der Entscheidungskriterien, und 3) Bewertung der Entscheidungsalternativen. Die Berechnung der Matrizen wurde mit Hilfe von MS Excel durchgeführt. Die einzelnen Schritte werden nachfolgend genauer beschrieben.

Schritt 1: Problemlösung und Festlegung der Zielhierarchie

Im ersten Schritt des Analytic Hierarchy Process wird zuerst das Entscheidungsproblem definiert und anschließend hierarchisch in globale und lokale Kriterien zerlegt. Wie in nachfolgender Abbildung (siehe: Abb. 9 nächste Seite) dargestellt, werden auf der untersten Ebene die Entscheidungsalternativen aufgeführt, darüber werden die Entscheidungskriterien eingeordnet.[49]

[48] Vgl. Riedl, R. (o. J.), S. 115 ff.
[49] Vgl. Saaty, T. (1999), S.407 ff.

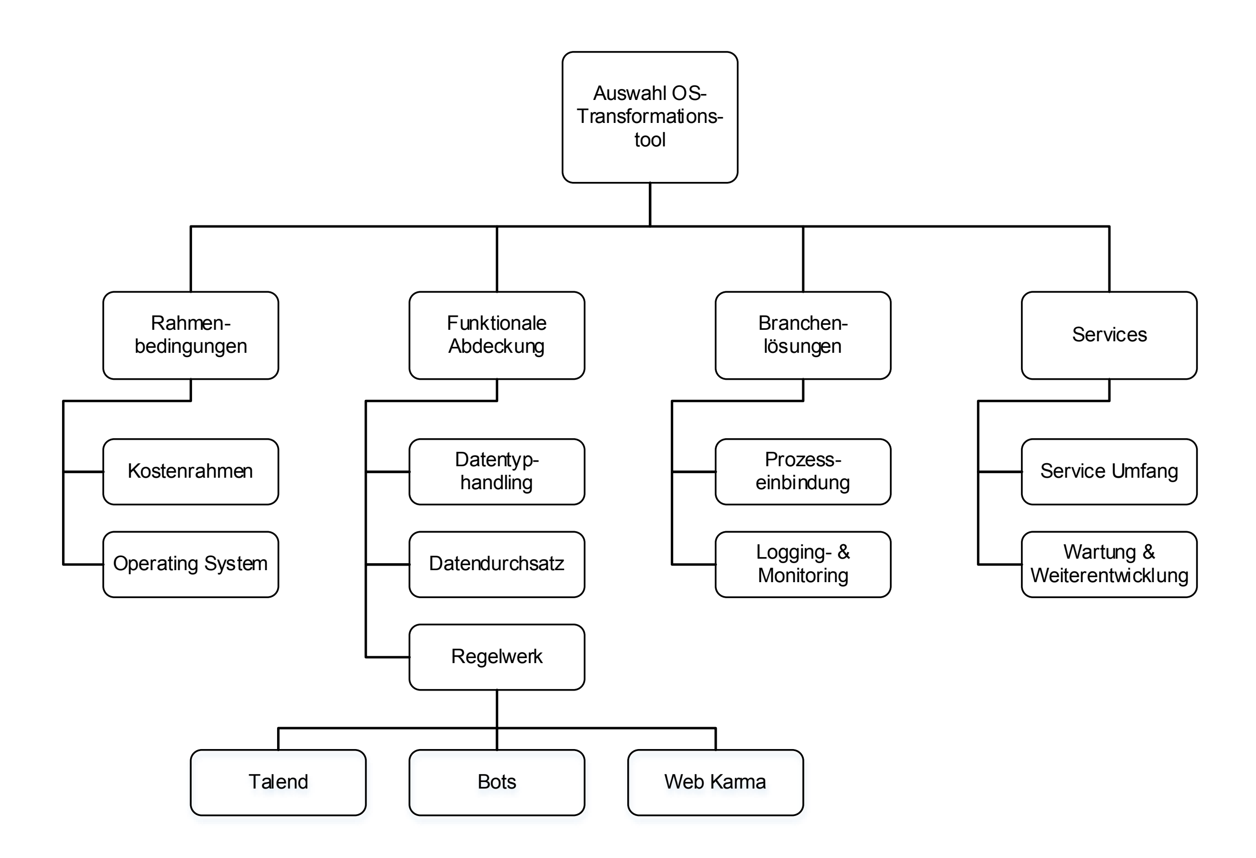


Abb. 9: Hierarchieebenen des AHP

Da der AHP die Ergebnisse der Nutzwertanalyse überprüfen und bestätigen soll, kann hier dasselbe Entscheidungsproblem verwendet werden. Ebenso werden die gleichen Kriterien für den Vergleich der Entscheidungsalternativen gewählt, welche auch zuvor in der Nutzwertanalyse angewendet worden sind. Die Auswahl der Entscheidungsalternativen richtet sich nach dem Ergebnis der vorhergegangenen Nutzwertanalyse. Da die Programme *Talend, Bots,* und *Web Karma* bei der Nutzwertanalyse die höchsten Bewertungen erzielt haben, werden diese drei Tools im Rahmen der AHP-Analyse erneut mit einander verglichen.

Schritt 2: Gewichtung der Entscheidungskriterien

Im zweiten Schritt des AHP werden die Entscheidungskriterien, welche derselben nächsthöheren Ebene zugeordnet sind, paarweise verglichen und gewichtet. Die Gewichtung wird basierend auf der eigens von Saaty entwickelten Skala (siehe Abb. 10) durchgeführt. Mit Hilfe von Matrizen werden die Paarvergleiche dargestellt und die Gewichtungsvektoren ausgerechnet, welche letztendlich die relative Gewichtung der Kriterien zeigen. Im Folgenden wird dieser Prozess genauer beschrieben.[50]

Bewertungsskala nach Thomas L. Saaty	
1	Gleiche Bedeutung
3	Etwas höhere Bedeutung
5	Sehr viel höhere Bedeutung
7	Erheblich höhere Bedeutung
9	Absolut dominierend
2, 4, 6, 8	Zwischenwerte

Abb. 10: Bewertungsskala nach Saaty[51]

Gewichtung der globalen Kriterien

Als erstes werden die übergeordneten Kriteriengruppen (auch: globale Kriterien) der höchsten Hierarchieebene, sprich die Gruppen „Rahmenbedingungen“, „Funktionale Abdeckung“, „Branchenlösungen“, und „Services“ verglichen und gewichtet. Dies erfolgt mit Hilfe der nachfolgenden Paarvergleichsmatrix (siehe Tab. 4). Dabei wird das Element einer jeden Zeile mit den Elementen aus jeder Spalte verglichen.

Gruppenbewertung:		(I)	(II)	(III)	(IV)	(ω)
(I)	Rahmenbedingungen	1	1/7	1/5	1/3	1/16
(II)	Funktionale Abdeckung	7	1	2	5	70/129
(III)	Branchenlösungen	5	1/2	1	5	5/17
(IV)	Services	3	1/5	1/5	1	3/34
	gesamt	16	1 59/70	3 2/5	11 1/3	

Tab. 4: Paarvergleichsmatrix der Kriteriengruppen

[50] Vgl. Riedl, R. (o. J.), S. 104
[51] Vgl. Saaty, T. (1999), S. 409

Am Beispiel der Kriteriengruppe „Services“ sieht man:

- Sie ist von erheblich geringerer Bedeutung als „Funktionale Abdeckung“ und „ Branchenlösungen“.
- Sie ist von etwas höherer Bedeutung als „Rahmenbedingungen“.

Nach Durchführung des Paarvergleichs erhält man die relative Gewichtung der globalen Kriterien, ausgedrückt durch den Gewichtungsvektor ω. Um zu garantieren, dass die obige Matrix widerspruchsfrei ist muss diese auf ihre Konsistenz geprüft werden. Die Konsistenzprüfung ist auch für den späteren Vergleich der einzelnen Kriterien und letztendlich für den Vergleich der Entscheidungsalternativen von hoher Bedeutung, da bei Missachtung das Ergebnis verfälscht werden kann.[52]

Konsistenzprüfung der Paarvergleichsmatrix

Bei der Konsistenzprüfung wird zunächst die Paarvergleichsmatrix mit dem Gewichtungsvektor ω multipliziert. Das Produkt dieser Berechnung hat wiederum den Vektor ω_s als Ergebnis.

$$\begin{pmatrix} 1 & 1/7 & 1/5 & 1/3 \\ 7 & 1 & 2 & 5 \\ 5 & 1/2 & 1 & 5 \\ 3 & 1/5 & 1/5 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1/16 \\ 70/129 \\ 5/17 \\ 3/34 \end{pmatrix} = \begin{pmatrix} 0.23 \\ 2.01 \\ 1.32 \\ 0.44 \end{pmatrix} = \omega_s$$

Anschließend wird aus dem Durchschnitt der Elemente des Produktes der Vektoren ω_s und $1/\omega$ der Eigenwert λ gebildet. Je näher sich der Eigenwert λ der Spaltenanzahl n der Paarvergleichsmatrix nähert (hier: n=4), desto konsistenter sind die Beurteilungen.

$$\omega_s \cdot \frac{1}{\omega} = \lambda = 4.22$$

Da in diesem Fall der Eigenwert $\lambda \neq n$ ist muss zunächst angenommen werden, dass es sich um eine inkonsistente Paarvergleichsmatrix handelt. Ob diese Inkonsistenz dennoch akzeptiert werden kann, lässt sich mit dem sogenannten Konsistenzindex C.I. (englisch: Consistency Index) und dem Zufallsindex R.I (englisch: Random Index) herausausfinden.
Während sich der Konsistenzindex mit nachfolgender Formel berechnen lässt, bemisst sich der Zufallsindex an der Dimension der Paarvergleichsmatrix.[53]

[52] Vgl. Riedl, R. (o. J.), S. 107 ff.
[53] Vgl. Riedl, R. (o. J.), S. 108

C.I. = $\frac{\lambda - n}{n-1}$

Für die zuvor genannte 4x4 Matrix ergibt sich daher ein R.I. = 0.9 [54] und ein C.I. = 0.07. Daraufhin lässt sich das relative Konsistenzverhältnis C.R. (englisch: Consistency Ratio) bestimmen, welches sich aus der Division des Konsistenzindexes C.I. und des Zufallsindexes R.I. zusammensetzt. Das Konsistenzverhältnis C.R. drückt letztendlich aus in welchem prozentualen Verhältnis die Paarvergleichsmatrix Resultat einer Zufallsverteilung ist. Generell gilt hier, dass C.R.-Werte < 0.1 trotz einer geringen Inkonsistenz akzeptiert werden können.
Bei C.R.-Werten >0.1 empfiehlt es sich die Beurteilungen zu überarbeiten.[55]

C.R. = $\frac{C.I.}{R.I.}$ = $\frac{0.07}{0.9}$ = 0.079

Da in diesem Fall der C.R.-Wert < 0.1 ist kann der Paarvergleich trotz einer geringen Inkonsistenz angenommen werden. Wie bereits zuvor beschrieben muss die Konsistenzprüfung auf alle Paarvergleiche, d.h. für alle Kriterien- und Produktvergleiche, angewendet werden, um eine widerspruchsfreie Auswahl zu garantieren. Auf weitere Konsistenzprüfungen wird hier allerdings nicht weiter eingegangen.

Gewichtung der lokalen Kriterien

Im weiteren Prozess werden nun die lokalen Kriterien innerhalb ihrer Kriteriengruppe analog zum vorigen Verfahren verglichen und bewertet. Die lokale Gewichtung der Kriterien wird im Anschluss zusammen mit der Gewichtung ihrer Gruppe zu einer globalen Bewertung verrechnet. Da die Gewichtung der Kriteriengruppen eine leichte Inkonsistenz aufweist, wird sich diese demnach auch auf die globalen Gewichtungen der Kriterien auswirken.

[54] Banai-Kashani, R. (o. J.), S. 689
[55] Vgl. Lamata, M./ Peláez, J. (2003), S. 1839 f.

Für die Kriterien im Rahmen der Auswahl des OS-Transformationstools ergeben sich folgende Gewichtungen, welche gerundet in Tab. 5 gezeigt werden.

Kriterium	Lokale Gewichtung	Globale Gewichtung
Kostenrahmen	0.5	0.03
Operating System	0.5	0.03
Datentyphandling	0.33	0.18
Datendurchsatz	0,33	0.18
Regelwerk	0.33	0.18
Prozesseinbindung	0.5	0.15
Logging und Monitoring	0.5	0.15
Serviceumfang	0.1	0.03
Wartung & Weiterentwicklung	0.9	0.26

Tab. 5: Gewichtung der Kriterien

Schritt 3: Gewichtung der Entscheidungsalternativen

Nachdem die Kriterien nun verglichen und gewichtet worden sind, werden die Merkmalsausprägungen der drei Transformationstools ebenfalls miteinander verglichen. Die lokalen Prioritäten der Entscheidungsalternativen werden erneut anhand der Saaty-Skala ermittelt.[56]

Für die Entscheidungsalternativen ergeben sich bezüglich des Merkmals „Datentyphandling" beispielsweise folgende Gewichtungen. Diese Gewichtungen werden auch in der Vergleichsmatrix zusammen mit dem lokalen Gewichtungsvektor ω dargestellt (siehe Tab. 6).

- Die Tools *Talend* und *Bots* haben bezogen auf das Datentyphandling eine gleiche oder sehr ähnliche Merkmalsausprägung. Sie werden daher gleich gewichtet.
- Da *Web Karma* keinen EDIFACT Standard unterstützt, erhält *Web Karma* im Vergleich zu *Talend* und *Bots* eine niedrigere Gewichtung bezüglich des Datentyphandling.

Datentyphandling					
Alternativen		(I)	(II)	(III)	ω
(I)	Talend	1	1	5	5/11
(II)	Bots	1	1	5	5/11
(III)	Web Karma	1/5	1/5	1	1/11
	Total	2.2	2.2	11	1

Tab. 6: Vergleichsmatrix "Datentyphandling"

[56] Vgl. Banai-Kashani, R. (o. J.), S. 690 f.

Sind für alle Alternativen die jeweiligen lokalen Prioritäten festgestellt, so werden diese mit den globalen Gewichtungen der Kriterien multipliziert um eine globale Priorisierung der Alternativen zu erhalten. Wie in Tab. 7 dargestellt, ergeben sich für die Auswahl der OS-Transformationstools folgende Gewichtungen. Es ist zu beachten, dass auf Grund der leichten Inkonsistenz der globalen Gewichtungen der Kriterien auch eine leichte Inkonsistenz bei der endgültigen Priorisierung der Entscheidungsalternativen vorzufinden ist.

Kriterium	**Globale Prioritäten**	**Alternativen**		
		Talend	**Bots**	**Web Karma**
Kostenrahmen	.03	0.33	0.33	0.33
Operating System	.03	0.43	0.43	0.14
Datentyphandling	.18	0.46	0.46	0.09
Datendurchsatz	.18	0.55	0.2	0.25
Regelwerk	.18	0.59	0.11	0.3
Prozesseinbindung	.15	0.33	0.33	0.33
Logging & Monitoring	.15	0.55	0.29	0.17
Serviceumfang	.03	0.4	0.4	0.2
Wartung & Weiterent.	.26	0.33	0.33	0.33
Summe	1.19	0.54	0.35	0.3
Verteilung in %		45%	30%	25%

Tab. 7: Priorisierung der Alternativen

5.2.4 Ergebnis der AHP-Analyse

Das Ergebnis der AHP-Analyse bestätigt die Reihenfolge der besten drei OS-Transformationstool aus der Nutzwertanalyse. Mit einer globalen Priorisierung von 0.54 geht *Talend* erneut mit Abstand als Sieger hervor, während die beiden Tools *Bots* und *Web Karma* eine sehr ähnliche Gewichtung haben. Wie berechnet kann die leichte Inkonsistenz der Beurteilungen missachtet werden und das Ergebnis akzeptiert werden.

5.3 Ergebnisse der Analysen

Zusammenfassend lässt sich sagen, dass von den ursprünglich sieben für die Analyse ausgewählten Open Source Datentransformationstools hauptsächlich *Talend*, *Bots* und *Web Karma* überzeugen konnten. Während diese drei Programme die Kriterien der Nutzwert –

und der AHP-Analyse gut oder sehr gut erfüllen konnten, schnitten die restlichen Tools in mindestens zwei oder mehreren Bewertungskriterien ungenügend oder mangelhaft ab. Im Rahmen der Analyse und Internetrecherche ist zudem aufgefallen, dass viele Transformationstools zwar mehrere Datentypen unterstützen, allerdings nur eine Transformation von oder zu XML zulassen. Komplexere Transformationen, beispielsweise von JSON zu EDIFACT, unterstützen nur die wenigsten dieser Programme. Weiterhin gibt es nur eine geringe Anzahl an Programmen, vor allem im Bereich der Open Source-Software, welche den EDIFACT Standard unterstützen. In unseren Tests konnten wir hierzu nur *Talend* und *Bots* ausfindig machen.

6 Vorstellung der Top 3 Produkte

Im folgenden Kapitel werden die drei Produkte: Bots, Talend und Web Karam vorgestellt. Diese kamen als Ergebnis der Nutzwert- bzw. AHP-Analyse heraus.

6.1 Bots

Bots ist ein „open source EDI translator"[57] unter der GNU GPL v3. Die zum Zeitpunkt des Projekts aktuelle Version ist 3.2.0, veröffentlicht am 2. September 2014.

Gründer und Haupt-Mitwirkender von Bots ist Henk-Jan Ebbers, auf der Projektseite werden noch zehn weitere Mitwirkende gelistet[58].

Ebbers bietet über seine Beratungsfirma EbbersConsult kommerziellen Support für Bots.[59]

Bots ist ein in der Programmiersprache Python (Version 2.7.x) geschriebenes Tool, das eine browserbasierte grafische Oberfläche hat. Dadurch kann es auf allen Betriebssystemen verwendet werden, für die es die Python-Umgebung gibt – Windows, Linux/Unix, Mac OS X sind die am häufigsten verwendeten Varianten.

Eine mögliche Verwendungsart ist somit die Installation auf einem Server im Intranet, wodurch eine zentrale Verwaltung ermöglicht wird.[60]

[57] Ebbers, H. (2014b)
[58] Vgl. Ebbers, H. (2014d)
[59] Vgl. Ebbers, H. (o. J.)
[60] Vgl. Ebbers, H. (2014e)

Support und Kommunikation finden über ein Wiki[61] und eine aktive Mailing-List (https://groups.google.com/forum/#!forum/botsmail) mit mehreren Beiträgen pro Tag statt.

6.1.1 Funktionalität

Datentypen:

Bots kann zwischen einer breiten Reihe an Formaten konvertieren. Diese können sogar noch individuell angepasst werden, um bestimmte Anforderungen zu erfüllen.

Folgende Formate sind möglich[62]:

- Edifact
- X12
- Tradacoms
- Xml
- Csv/delimited
- Fixed
- Excel
- Json
- SAP idoc
- Eancom
- Html
- Direct database communication

Diese Daten können auf verschiedene Art zur Verfügung gestellt werden:[63]

- Dateisystem
- E-Mail-Protokolle (Pop3, Imap, SMTP)
- Verschiedene FTP-Protokolle
- XML-RPC
- http(s)
- Datenbank-Verbindungen
- Benutzerdefinierte Verbindungen

Fehlerbehandlung und Logging:

Falls Fehler bei einer Umwandlung auftreten, werden diese sowohl in der Benutzeroberfläche im Browser angezeigt als auch in eine Logdatei geschrieben. In den Konfigurationsdateien kann der Detailgrad des Logs eingestellt werden. So können beispielsweise zum Debugging ausführlichere Informationen angezeigt werden.

[61] Vgl. Ebbers, H. (2014f)
[62] Vgl. Ebbers, H. (2014a)
[63] Vgl. Ebbers, H. (2014a)

6.1.2 Ablauf der Transformation

Die Umwandlung von einem Datenformat in ein anderes besteht aus mehreren Komponenten. Im Folgenden wird kurz der Ablauf einer Transformation in Bots beschrieben, Grundlagen dafür bietet das Wiki auf der Projektseite.[64] (Siehe auch Abb. 11 auf der folgenden Seite)

Channel (Kanäle) beschreiben die Art, wie Dateien in das System eingelesen und am Ende ausgegeben werden. Wenn das Dateisystem die Quelle ist, werden hier Dateiname und Verzeichnis hinterlegt, bei Netzwerkquellen Host, Benutzername und Passwort.

Routen bestimmen zwei Channel, Input und Output, sowie Datentyp und Grammatik (Grammar) des Inputs.

Da es für jeden Datentyp unterschiedliche Formate geben kann, bestimmen die Grammatiken, welche Inhalte an welcher Stelle in den Daten stehen. Sie legen somit die Syntax der Dateien fest. Diese besteht aus der Länge der einzelnen Felder, ihrer Position in der Datei, ihrem Datentyp (String, Integer, Decimal, Datum) und einem Booleschen Wert, ob dieses Feld verpflichtend oder optional ist.

Translations (Übersetzungen) sind die Kernverbindungen zwischen Input und Output. Durch sie wird festgelegt, wie eine eingehende Datei in einem bestimmten Format in das gewünschte Zielformat umgewandelt wird. Für jede mögliche Transformation werden eingehender Datentyp, eingehende Grammatik, ausgehender Datentyp und ausgehende Grammatik bestimmt.

Ebenso wird in den Translations auf ein Mapping Script (Zuordnungsskript) verwiesen, in dem die konkrete Zuordnung der Input-Felder zu den Output-Feldern vorgenommen wird. In diesem Skript werden darüber hinaus alle Änderungen an den Daten definiert. So kann hier beispielsweise ein Datenfeld für das Geschlecht von „1“ und „2“ zu „m“ und „w“ geändert oder Felder kombiniert und getrennt werden.

[64] Vgl. Ebbers, H. (2014c)

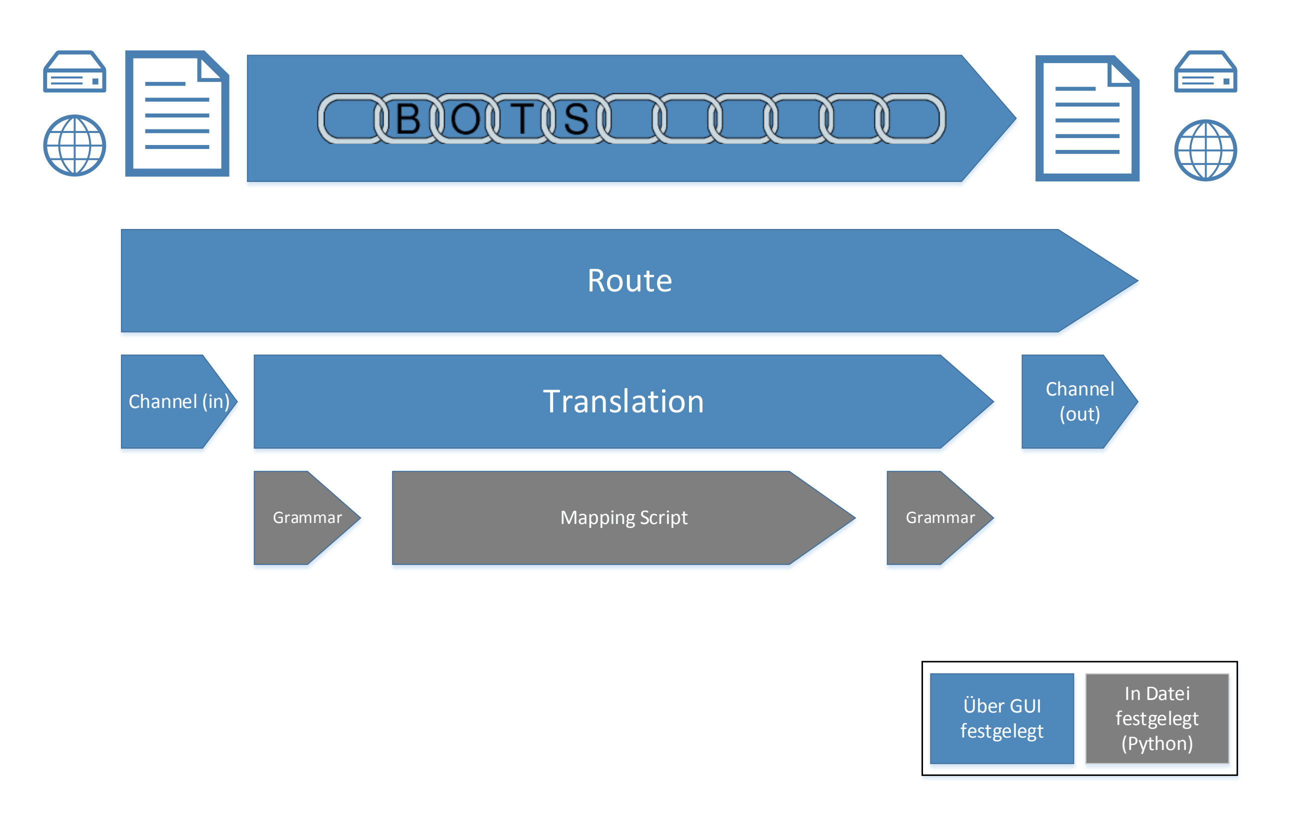


Abb. 11: Modell des Ablaufs einer Transformation bei Bots

6.2 Talend

Talend wurde 2006 von Fabrice Bonan in Frankreich gegründet. Heute hat es einen zweigeteilten Firmensitz in Redwood City, Kalifornien und Suresnes, Frankreich. Die hauptsächliche Entwicklung der Produkte findet in China statt. Mit 400 Mitarbeitern erzielen sie ein Durchschnittliches Umsatzwachstum von 108% pro Jahr. In Deutschland haben sie einen Standort in Bonn. Ihren Internetauftritt und eine aktive Community findet man unter https://de.talend.com/.

Um ihre Produkte stetig weiter zu entwickeln und sie auf möglichst vielen Plattformen anbieten zu können, arbeitet Talend mit über 30 Partner eng zusammen. Darunter sind bekannte Unternehmen wie Google.

Zu Talends Angeboten gehört OS Software, die in erster Linie für Unternehmen konstruiert ist und auf datenzentrierten Geschäftsmodelle basiert. Dabei wird die Möglichkeit geboten, einfach und schnell auf verschiedene Datentypen zugreifen zu können. Talend unterstützt sowohl ältere Datenformate als auch aktuelle. Durch Umsetzung einer Zero-Footprint-Lösung ohne feste Vorgaben der Datenstrukturen wird ebenso ein Lösungsansatz für komplett neue Datenquellen angeboten. Dies ist besonderer Vorteil im Bereich Big Data, in dem sich Datenquellen ständig verändern und weiterentwickeln.

Big Data ist der Hauptbereich, auf den sich Talend konzentriert. Neben dieser Lösung für Big Data wird eine Open Source Umgebung angeboten, in der es möglich ist, auf eine einfache Art und Weise Daten zu extrahieren, zu transformieren und diese in einem umfangreichen Maß zu verarbeiten. Des Weiteren bietet Talend sieben Plattformen für verschiedene Anforderungen und neben fünf weiteren Tools noch diverse andere Lösungen an. Die bereitgestellten Tools sind:

- Business Process Management: Tool zur Modellierung, Erstellung und Optimierung von Geschäftsprozessen

- Data Integration: Tool, mit dem sich Daten aus verschiedenen Geschäftsanwendungen, sowohl in Echtzeit als auch im Batch Modus, umwandeln und integrieren lassen. Hierbei werden sowohl operative, als auch analytische Anforderungen der Datenintegration berücksichtigt.

- Data Quality: Dieses Tool beinhaltet ein End-to-End Profiling sowie eine Überwachungsfunktion von Daten, um Unstimmigkeiten zu identifizieren. Dabei werden auch Zusammenhänge zwischen Daten erkannt, um die Datenqualität der Anwendungen zu gewährleisten.

- Enterprise Service Bus: Mit Hilfe dieses Tools wird die Erstellung, Verbindung, Vermittlung und Verwaltung von Services und deren Interaktion deutlich vereinfacht.

- Master Data Management: Die Hauptaufgabe dieses Tools besteht darin, eine zentrale Plattform für die Stammdatenverwaltung bereitzustellen. Daneben bietet es auch Daten Stewardship, Datenintegration und Datenqualität Funktionen.

Alle Produkte, die von Talend angeboten werden, basieren auf einer Weiterentwicklung von Eclipse und der Programmiersprache Java. Entsprechend der Anforderungen wurde im Rahmen dieser Arbeit mit dem Programm Talend Data Integration gearbeitet.

Talend läuft unter der im Kapitel 2.2.1 beschriebenen Apache-Lizenz v2. Dies hat für das Projekt verschiedene Vor- und Nachteile. Zu den Vorteilen zählt, dass Produkte, die unter der Apache-Lizenz laufen, von jedem kostenfrei genutzt und an persönliche Bedürfnisse angepasst werden können. Diese Weiterentwicklungen dürfen dann unter einer beliebigen Lizenz wieder veröffentlicht werden. Das bedeutet, dass die so entwickelten Produkte auch kommerziell vertrieben werden können. Dies stellt sowohl einen Vor- als auch einen Nachteil

dar, da Weiterentwicklungen oft nicht veröffentlicht werden und so viel an Innovativen Entwicklungen zugekauft oder gar nicht verfügbar sind.[65]

6.2.1 Funktionalitäten

Datentypen:

Talend Data Integration stellt verschiedene vordefinierte Schnittstellen bereit mit denen diverse Datentypen konvertiert werden können. Neben den vordefinierten Schnittstellen gibt es auch die Möglichkeit die Formate individuell anzupassen, um auch neuere Datentypen verarbeiten zu können. Zu den Datentypen, für die Schnittstellen bereits vorhanden sind, gehören: Edifact, XML, CSV (generell delimited Daten), Excel, JSON, HTML, REGEX, Idif, LDAP, Salesforce, GDV (generell positional Datentypen) sowie Direct Database Communication.

Die Einbindung der Dateien kann durch ein Dateisystem, eine http(s) Anbindung und durch eine Datenbankanbindung geschehen.

Fehlerbehandlung und Logging:

Talend ist in der Lage, Fehler und Problem in der Konsole auszugeben. Ein eigenes Fehlerhandling mit Logdateien sowie Monitoring ist in der OS-Version nicht enthalten. Diese Funktionen können durch Erwerb des Talend Enterprise Data Integration Programms hinzugekauft werden.[66]

6.2.2 Ablauf der Transformation

Der Vorgang der Transformation besteht aus drei bis fünf Elementen. Im Folgenden werden diese Elemente beschrieben, basierend auf der bereitgestellten Dokumentation und eigenen Erfahrungen.[67]

Talend stellt eine grafische Oberfläche auf der die wesentlichen Änderungen mit einfachen Drag&Drop-Aktionen durchgeführt werden können. Nur tiefergehende Spezifikationen kön-

[65] Vgl. Die Beauftragte der Bundesregierung für Informationstechnik, Bundesministerium des Innern (2012), S. 105 ff.
[66] Vgl. Talend Germany GmbH (**o. J.a):**
[67] Vgl. Talend Germany GmbH (**o. J.b):**

nen durch JAVA Befehle spezifiziert werden. Zu diesem Zweck hat jedes Element ein eigenes Tab, in dem die Spezifikationen durchgeführt werden können.

Der grundsätzliche Vorgang einer Konvertierung besteht aus mindestens drei Elementen, einem Input-Element, einer Map sowie einem Output-Element.

Mit Hilfe des Input-Elements kann die Quelldatei geladen werden. Dies geschieht durch die Auswahl einer dem Datentyp entsprechenden Schnittstelle sowie dem Festlegen des Verzeichnisses und des Dateinamens, von der die zu konvertierende Datei bezogen werden soll. Dieses Element stellt den Beginn der Transformation dar.

Die Map wird als Hauptverbindung an das Input Element angehängt. Sie ermöglicht es, ein Regelwerk zu implementieren. Hier kann sowohl das Input- als auch das Output-Schema definiert werden sowie jede Spezifikation, die während der Transformation durchgeführt werden soll, um die Datensätze von dem einen Schema in das andere zu konvertieren. Dieses Element bietet bereits vorgefertigte Transformationsregeln an, allerdings können auch vollständig eigene Regeln definiert werden. Dies geschieht durch einfache Javabefehle, bzw. Drag&Drop.

Abschließend wird die Hauptverbindung an das Output-Element angehängt. Hier wird festgelegt in welchen anderen Dateityp konvertiert werden soll. Dies geschieht erneut durch die Auswahl einer entsprechenden Schnittstelle. Hier kann des Weiteren ein beliebiger Seicherort ausgewählt werden.

Diese drei Elemente sind das Grundgerüst jeder Dateitransformation und erlauben, es einzelne Dateien zu konvertieren. Um alle im Rahmen dieser Arbeit geforderten Anforderungen zu erfüllen, wie die Konvertierung von N Dateien des Typs A zu N Dateien des Typs B, werden noch zwei weitere Elemente benötigt. Das ist zum einen das Element tFileList und eine weitere Input Datei des Typen Delimited. Dies wird konkreter in der Dokumentation beschrieben und führt an dieser Stelle zu weit.

6.3 Web Karma

Web Karma, oder auch nur Karma genannt, ist ein Datenintegrationstool, welches von der University of Southern California entwickelt wurde. Da von dieser Software im Zuge dieser Arbeit kein Prototyp entwickelt wurde, konnte keine vergleichbar ausführliche Beschreibung der Software erstellt werden. Zum Zeitpunkt des Projekts ist die aktuelle Version von Web

Karma v2.033, welche am 24. November 2014 veröffentlicht wurde. Da Web Karma unter die Apache 2 Lizenz fällt, handelt es sich wie Bots und Talend um Open Source Software.[68]

Web Karma kann über die Plattform *GitHub* bezogen werden, auf welcher sowohl der Support als auch die Kommunikation abläuft. Dort kann auf ein Wiki und ein aktives Forum zurückgegriffen werden.

Web Karma ist in der Programmiersprache Java geschrieben und kann daher auf Mac, Linux und Windows eingesetzt werden. Generell handelt es sich bei Web Karma um eine webbasierte Applikation, d.h. die graphische Oberfläche wird mittels einem Browser dargestellt. Normalerweise laufen Server und Client auf einer Maschine, Web Karma kann allerdings auch auf einem Server installiert werden und von unterschiedlichen Geräten genutzt werden. [69]

Web Karma unterstützt eine Vielzahl unterschiedlicher Datentypen und Datenquellen für die Konvertierung. Die Transformation lässt beispielsweise Spreadsheets, XML, JSON, oder sogenannte „delimited text files", wie z.B. CSV oder GDV, zu. Die Daten können dabei manuell eingepflegt werden, direkt von relationalen Datenbanken (MySQL, SQL Server, Oracle und PostGIS) importiert werden, oder von Web APIs geladen werden. Weiterhin können hohe Datenmengen an JSON-, XML-, CSV- und Datenbank-Dateien im Batch Mode verarbeitet und entweder zu RDF oder JSON konvertiert werden. [70]

7 Prototyperstellung

Auf Basis der Ergebnisse aus der Nutzwert- und AHP-Analyse aus Kapitel 5 werden von den zwei Produkten Bots und Talend zwei Prototypen mit identischem Funktionsumfang erstellt.

7.1 Prototyp Bots

Vorgehen bei der Implementierung von Bots:

- Installation und Start des Programms und Einstellungen
- Erstellung von Grammars und Mapping Scripts

[68] Vgl. Information Sciences Institute, University of Southern California (o. J.)
[69] Vgl. o. V. (**o. J.b):**
[70] Vgl. o. V. (**o. J.c):**

- Erstellung von Channels, Routes, Translations
- Starten von Transformationen
- Erstellung der Testumgebung

7.1.1 Installation und erste Schritte

Bots benötigt Python in der Version 2.7.x (https://www.python.org/downloads/) sowie die Bibliotheken Django in einer Version zwischen 1.4.0 und 1.7.0 (https://www.djangoproject.com/download/) und cherrypy in Version 3.1.0 oder neuer (https://pypi.python.org/pypi/CherryPy).[71] Nachdem diese Voraussetzungen installiert sind, kann Bots heruntergeladen(http://sourceforge.net/projects/bots/files/bots%20open%20source%20edi%20software/3.2.0/) und installiert werden. Es speichert sich dabei im Ordner /python27/Lib/site-packages/bots/.

Die grafische Oberfläche des Tools wird durch das Aufrufen der Datei bots-webserver.py im Ordner /python27/Scripts/ gestartet. Standardmäßig kann die GUI im Browser unter http://localhost:8080/ aufgerufen werden. Als Benutzername und Passwort sind „bots“ bzw. „botsbots“ hinterlegt.

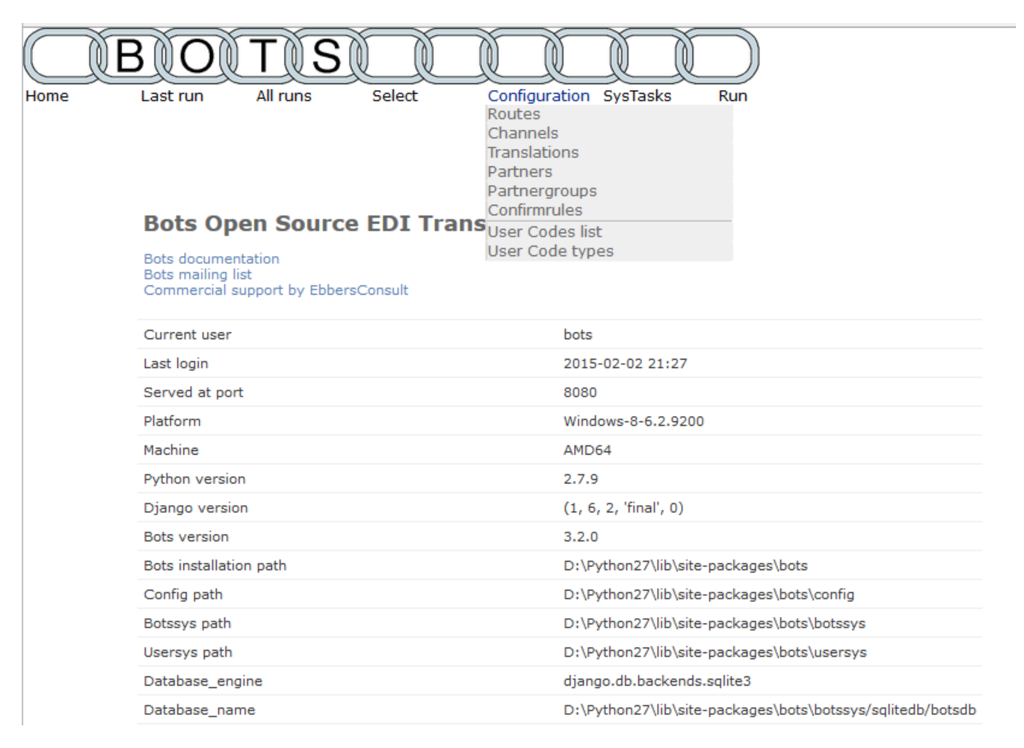


Abb. 12: Grafische Oberfläche von Bots

[71] Vgl. Ebbers, H. (2012a)

Über die grafische Oberfläche (siehe Abb. 12) können unter anderem Routes, Channels und Translations erstellt, Transformationsläufe gestartet sowie Reports der letzten Läufe betrachtet werden. Ebenso können hier eine Sicherung aller Einstellungen erstellt und Benutzer verwaltet werden. Weniger häufig verwendete Einstellungen werden in den Konfigurationsdateien /bots/config/bots.ini und /bots/config/settings.py vorgenommen.

7.1.2 Erstellung von Grammars und Mapping Scripts

Jede Transformation benötigt zwei Grammars, die die Struktur von Input und Output vorgeben, sowie ein Mapping-Script, das die Datentransformation durchführt. Diese Anweisungen werden in eigene Dateien in nativem Python-Code geschrieben – Zuordnungen finden über eine Schachtelung von Dictionaries und Lists, zwei in Python verwendeten Datenstrukturen, statt. Die Dateien werden unter /bots/usersys/grammars/{edi-type}/ bzw. /bots/usersys/mappings/{edi-type}/ gespeichert.

Grammar-Dateien haben drei Hauptbestandteile:[72]

- Ein Dictionary „syntax“, in dem generelle Einstellungen wie Zeichensatz und Trennzeichen vorgegeben werden
- Eine Liste „structure“, die die hierarchische Struktur der Datei vorgibt.
- Ein Dictionary „recorddefs“, das einzelne Datenfelder ihrem Vaterelement. Dort wird darüber hinaus festgelegt, ob das Feld verpflichtend oder freiwillig ist, welche Länge es hat und welcher Datentyp es ist

In einem Mapping-Script werden Daten aus der eingehenden Grammar aufgerufen und in das von der ausgehenden Grammar benötigte Format umgewandelt. Die Zuordnung erfolgt gemäß der in den Grammars festgelegten Hierarchie. In dieser Datei können alle Modifikationen hinterlegt werden, die an den Rohdaten vorgenommen werden sollen, da normaler Python-Code verwendet werden kann.

Eine mögliche Anweisung ist:

```
out.put({„BOTSID“: „root“}, {„foo“: inn.get({„BOTSID“: “source“}, {„bar“: None})})
```

[72] Vgl. Ebbers, H. (2012b)

Diese bewirkt, dass das Element „bar“ unter dem Wurzelknoten „source“ des Inputs gelesen und in das Feld „foo“ im Wurzelknoten „root“ geschrieben wird.

7.1.3 Erstellung von Channels, Routes und Translations

Channels, Routes und Translations binden diese Dateien im Ablauf der Transformationen ein. Im Channel wird bestimmt, woher und wohin die Daten geschrieben werden sollen. Bei Festlegung Abruf bzw. Speicherung als Datei wird der Pfad im Dateisystem sowie der Dateiname hinterlegt. Channels werden ganz am Anfang und am Ende der Transformation aufge-

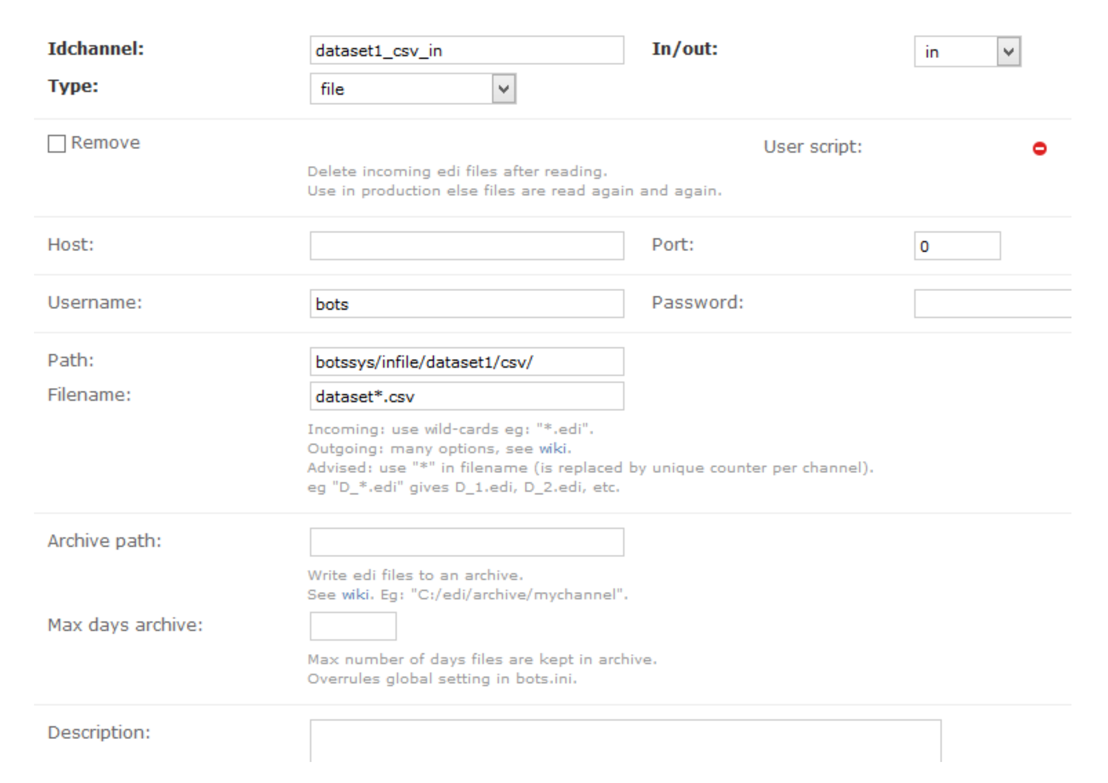

Idchannel: dataset1_csv_in **In/out:** in
Type: file
Remove
Delete incoming edi files after reading.
Use in production else files are read again and again.
User script:
Host: Port: 0
Username: bots Password:
Path: botssys/infile/dataset1/csv/
Filename: dataset*.csv
Incoming: use wild-cards eg: "*.edi".
Outgoing: many options, see wiki.
Advised: use "*" in filename (is replaced by unique counter per channel).
eg "D_*.edi" gives D_1.edi, D_2.edi, etc.
Archive path:
Write edi files to an archive.
See wiki. Eg: "C:/edi/archive/mychannel".
Max days archive:
Max number of days files are kept in archive.
Overrules global setting in bots.ini.
Description:

Abb. 13: Erstellung eines Channels

rufen, für jede Umwandlung werden also zwei Channel benötigt.

Translations legen das Mapping-Script fest, das zur Konvertierung von der Nachrichtenart (also Grammars) des Inputs zur Nachrichtenart des Outputs benötigt wird. Eine Translation hat daher fünf Felder (siehe Abb. 14): Eingehender EDI-Datentyp, eingehende Grammar,

Fromeditype: csv — Editype to translate from. **Frommessagetype:** dataset1_gdv — Messagetype to translate from.
Mapping Script: dataset1_to_gdv — Mappingscript to use in translation.
Toeditype: fixed — Editype to translate to. **Tomessagetype:** dataset1_csv — Messagetype to translate to.
Description:

Abb. 14: Erstellung einer Translation

das zu verwendende Mapping-Script, ausgehende Grammar und ausgehender Datentyp.

In der Route werden Channel und Translations verknüpft. Zu diesem Zweck benötigt eine Route vier Angaben (siehe Abb. 15): Eingehender Channel, eingehender EDI-Datentyp, eingehende Grammar und ausgehender Channel. Zudem wird hier festgelegt, ob eine Transformation stattfinden soll oder nur eine Bewegung der Daten von eingehendem zu ausgehendem Channel. Wenn eine Transformation vorgenommen werden soll, wird die benötigte

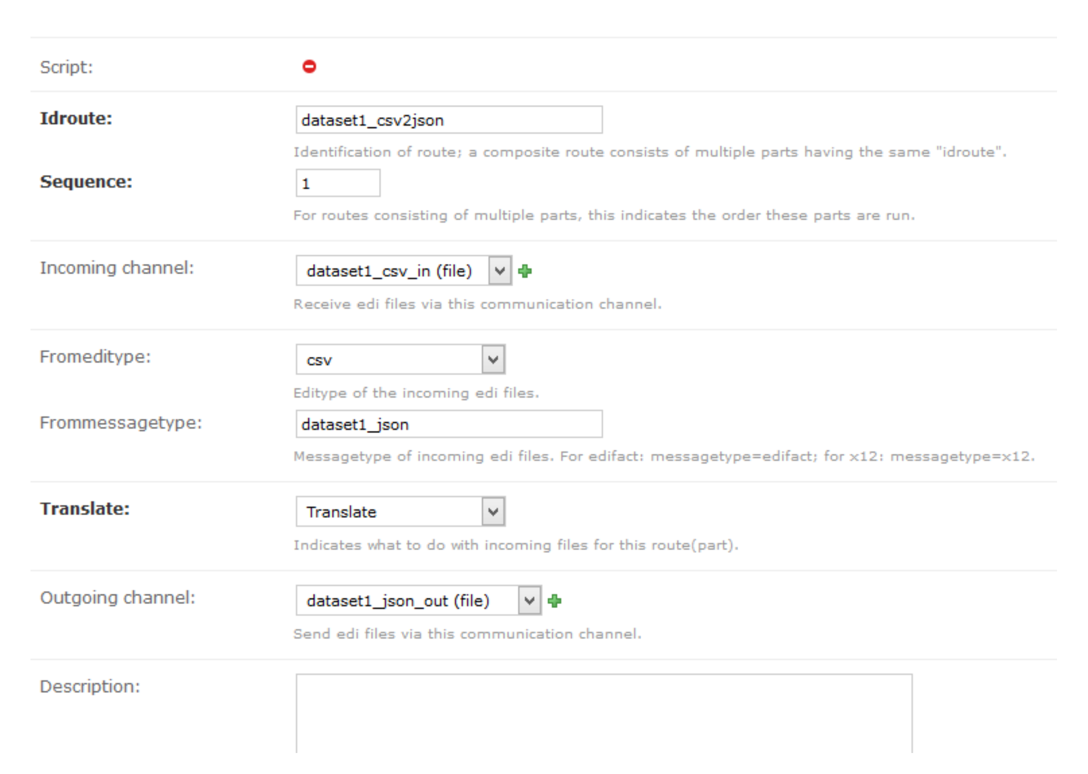


Abb. 15: Erstellung einer Route

Translation automatisch anhand der eingehenden Grammar gefunden.

7.1.4 Starten einer Transformation

Zum Starten einer Transformation gibt es mehrere Möglichkeiten.[73] Zum einen kann eine Route zu bestimmten Zeiten mithilfe eines cronjobs in Linux/Unix oder des Windows Task Schedulers gestartet werden. Dies ist nützlich, wenn Transformationen nicht bei Bedarf, sondern in regelmäßigen Abständen durchgeführt werden sollen. Dazu muss über das gewünschte Planungstool nur die Datei /python27/Scripts/bots-engine.py aufgerufen werden.

[73] Vgl. Ebbers, H. (2012c)

Über den Parameter „--new“ können alle aktivierten Routen gestartet werden, bei Nennung einer bestimmten Route nur diese.

Alternativ kann über die Aktivierung eines Directory Monitors, einer Verzeichnisüberwachung, gestartet werden. Bots wird dabei dann gestartet, wenn neue Dateien in bestimmten lokalen Verzeichnissen vorhanden sind. Diese Funktion ist vor allem dann nützlich, wenn Daten zu unvorhersehbaren Zeiten konvertiert werden müssen.

Zu guter Letzt können Transformationen auch manuell über die grafische Oberfläche im Browser gestartet werden.

7.1.5 Konfiguration der Umgebung für den Funktions- und Durchsatztest

Im anschließenden Durchsatz- und Funktionstest wurde auf das Testen mit Edifact-Datensätzen verzichtet, da diese sehr stark branchenspezifisch sind. Dadurch war es den Verfassern nicht möglich, ohne zur Verfügung gestellte Testdaten einen relevanten Test durchzuführen.

Um die Benutzung von Bots in den Tests zu erleichtern, wurde für jeden der in den Tests verwendeten Datentypen zwei Channel, Input und Output, erstellt. Jede der 24 getesteten Kombinationen (vier Datentypen in den Ausführungen direkte Transformation und Verwendung eines Regelwerks) bestand darüber hinaus aus einer eigenen Route sowie einem eigenen Mapping Script. Bei der Verwendung des Regelwerks werden für Input und Output in einem Format unterschiedliche Grammars benötigt, da durch die Regeln andere Felddefinitionen erforderlich sind. Daher sind im Test insgesamt 36 Grammars (neun pro Datentyp) verwendet.

Alle Durchläufe fanden getrennt voneinander statt, um Einflüsse auf die Geschwindigkeit zu minimieren. Sämtliche temporären Daten sowie alter Output wurden nach jedem Vorgang gelöscht.

7.2 Prototyp Talend

7.2.1 Installation und erste Schritte

Die Installationsdatei für Talent Data Integration kann auf deren Webseite unter http://talend.com/download heruntergeladen werden. Voraussetzung für die Installation ist Java der Version 1.7. Mit höheren Versionen ist Talent zurzeit nicht kompatibel. Des Weite-

ren wird das JAVA JDK benötigt, welches unter: http://www.oracle.com/technetwork/articles/javase/index-jsp-138363.html herunter geladen werden kann.

Im Workspace finden alle Arbeiten während des Projekts statt. Es gibt vier Bereiche die in der folgenden Abbildung (Abb. 16) gekennzeichnet sind.

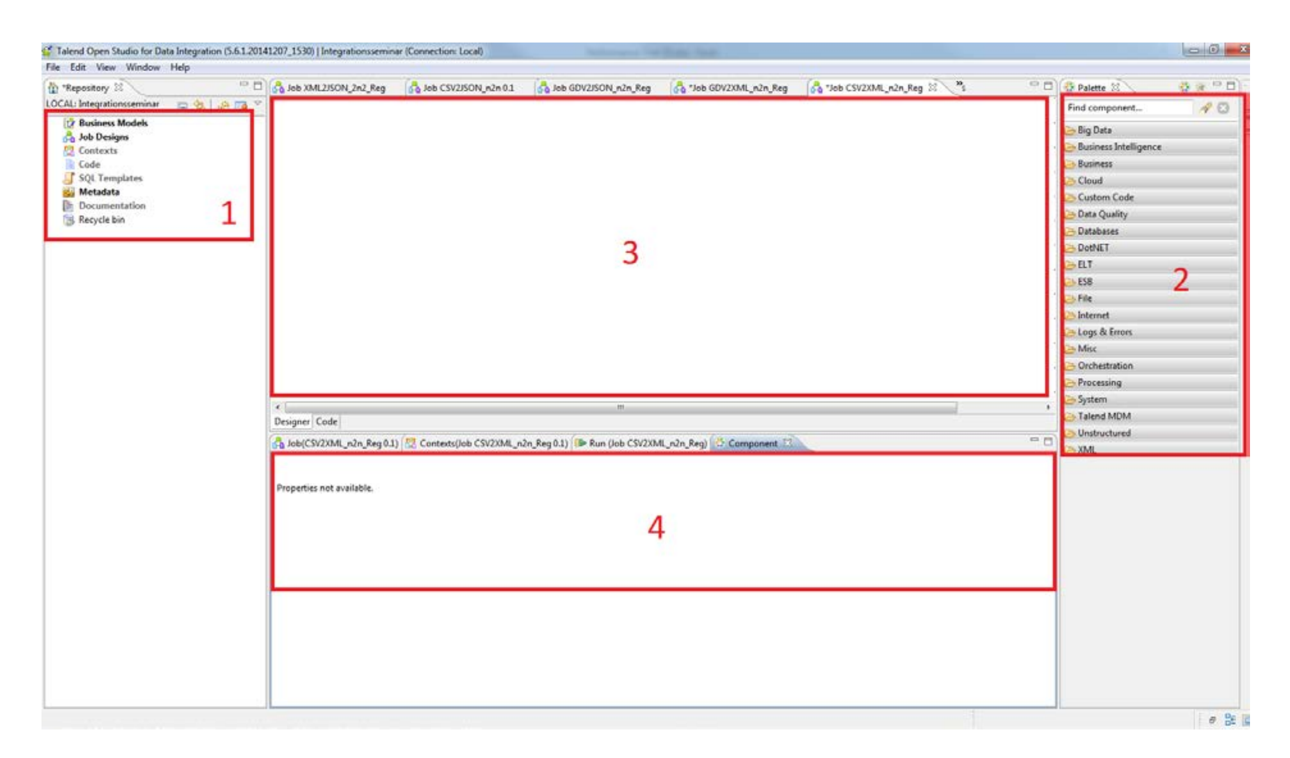


Abb. 16: Benutzeroberfläche Talend

Im Bereich auf der linken Seite (1) findet man das Repository. Hier werden alle Jobs gespeichert und importierte Dateien unter Metadaten verwaltet. Auf der rechten Seite (2) in der Palette sind alle Elemente, die man nutzen kann aufgeführt. Diese können mit einem einfachen Herüberziehen in den mittleren Bereich (3) genutzt werden. Dort findet auch die Definierung der Jobs statt. Im unteren Bereich (4) ist die Konsole und auch andere Tabs mit denen man spezifischere Einstellungen vornehmen kann.

Für einen Job sind verschiedene Elemente nötig. Dazu gehören ein Input, ein Output und ein Mapping Element. Damit der Input auf den Output gemapped werden kann, müssen beide ein Schema haben.

Ein solches Schema kann entweder erstellt werden, indem man es von einer bereits bestehenden Datei kopiert und es importiert oder ein eigenes Schema in den Komponenteneinstellungen erstellt. Schemas haben drei Aufgaben.

- Strukturierung: Bestimmt die Hierarchie zwischen den Daten
- Spezifizierung der Felder: Gibt an welchen Typ ein Feld hat, wie lang der Zeichensatz sein darf und ob Werte gesetzt sein müssen.
- Vorgabe der Syntax: Legt den Zeichensatz und die Trennzeichen einer Datei fest.

7.2.2 Erstellung eines Jobs

Exemplarisch wird die Transformation von CSV zu XML beschrieben. Anschließend werden weitere Elemente beschrieben, die als Schnittstelle für In- und Output verwendet werden können. Dabei wird besonders auf zu dem Beispiel abweichenden Konfigurationen eingegangen.

Für das Einlesen einer CSV Datei wir das Element tFileInputDelimited benötigt. Diesem Element muss entsprechend der zu konvertierenden Datei ein Schema vorgegeben werden, sodass es die Datei lesen kann. Um eine spezielle Datei zu verwenden, muss auch der Pfad zu der Datei vorgegeben werden. Beides kann in den Component-Einstellungen vorgenommen werden.

Um den Output als XML zu erhalten wird das Element tAdvancedFileOutputXML benötigt. Hier müssen ebenfalls Schema und Pfad definiert werden, wo die Datei gespeichert werden soll. Bei XML ist zu beachten, dass durch ein Doppelklick auf das Element die XML-Baumstruktur noch zu bearbeiten und das Loop Element zu setzen ist.

Das Schema des Inputs und des Outputs müssen nicht unbedingt gleich sein. Daher gibt es die Möglichkeit, durch das Element tMap Transformationsregeln zu definieren. Exemplarisch sieht man in Abb. 17 Transformationsregeln. Dabei werden Zeilen zusammengelegt, N1-N3 sind im Output-Schema nur noch eine Zeile, oder aus dem Feld GDat (ein Datum) werden drei Zeilen für jeweils Tag, Monat und Jahr.

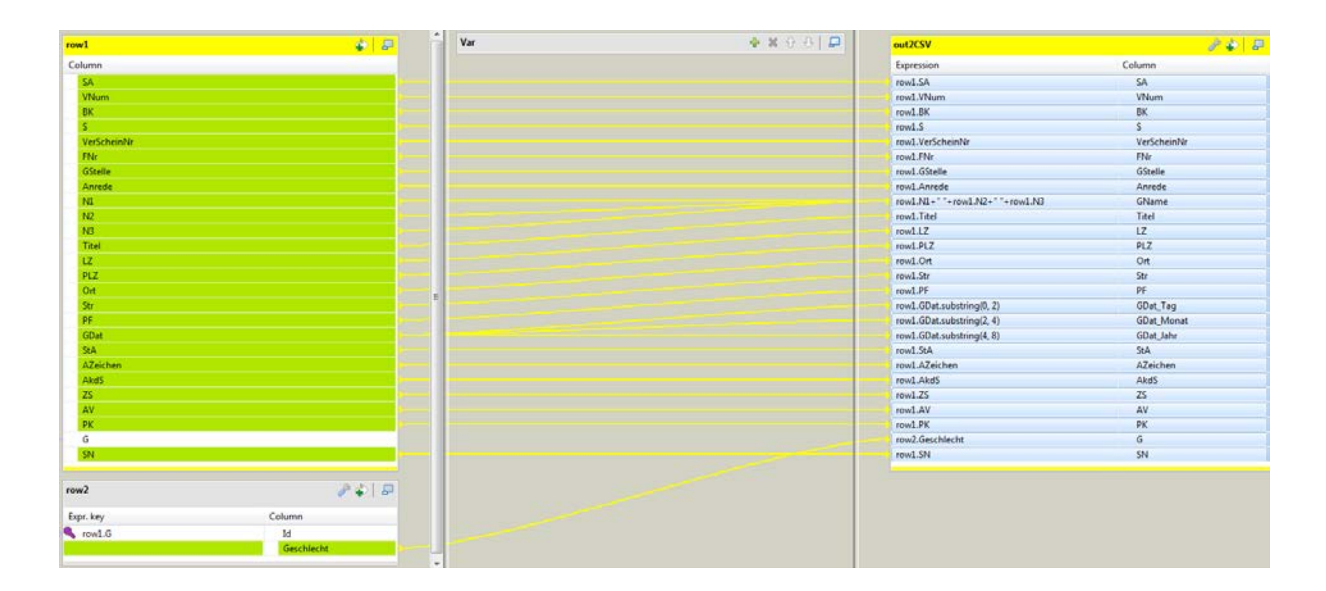


Abb. 17: Definition der Transformationsregeln in Talend

Nach dem Mapping kann durch Drücken von F6 der Job ausgeführt werden und die Transformation wird durchgeführt.

In dem beschriebenen Beispiel wurde nun eine Datei in eine andere konvertiert. Soll mehr als eine Datei konvertiert werden, muss das Beispiel angepasst werden.

Der wichtigste Punkt für die Konvertierung mehrerer Dateien ist, dass das Element tFileList hinzugefügt wird und mit dem Input-Element iterativ verknüpft wird. Nun kann in den Component-Einstellungen der tFileList der Ordner ausgewählt werden, aus dem die Dateien gelesen werden sollen. Des Weiteren muss der Pfad des Input- und des Output-Elements angepasst werden.

Für den Pfad der Input-Datei muss das folgende Statement eingetragen werden. Dieser String sorgt dafür, dass der Pfad des Input-Elements immer auf dem von tFileList aktuell ausgewählter Datei liegt.

```
((String) globalMap.get ("tFileList_1_CURRENT_FILEPATH"))
```

Ebenso muss der Output-Pfad angepasst werden, da sonst immer die gleiche Datei überschrieben würde, oder aber alles in eine Datei gespeichert wird.

Um den Ausdruck hier möglichst kurz zu halten wird /*/ als Platzhalter für ((String) globalMap.get("tFileList_1_CURRENT_FILE")) verwendet

/*Pfad zu gewünschtem Speicherort*/ + /*/.substring(0, /*/.length()-4)+ ".gewünschte Dateiendung"

7.2.3 Component Einstellung

In diesem Bereich (siehe Tab. 8) werden die Schnittstellen genannt und kurz beschrieben welche Aspekte zu konfigurieren sind wenn, sie genutzt werden.

Datei Typ	Verwendetes Element	Anforderungen
CSV	Input: tFileInputDelimited Output: tFileOutputDelimited	Schema
GDV	Input: tFileInputPositional Output: tFileOutputPositional	Schema + Spezifikation mit Zeichenlän-ge
JSON	Input: DataSetJSON Output: tFileOutputJSON	Schema Loop JSON Path query
XML	Input: DataSetXML Output: tAdvancedFileOutputXML	Schema Loop XPath query

Tab. 8: Elemente für In- und Output

7.2.4 Transformationsregeln

In diesem Abschnitt wird erläutert, wie die in den Tests verwendeten Transformationsregeln umgesetzt wurden.

<u>Regel 1 – Ersetzen von Werten</u>

Für die erste Regel sollten die Werte 1, bzw. 2 zu „männlich" bzw. „weiblich" transformiert werden. Dies wurde durch eine Lookup-Funktion, die von der tMap bereitgestellt wird, umgesetzt. Konkret wurde ein zweites Input-Element definiert, TReg, das die Zuweisung enthielt das die Id 1 für „männlich" steht und 2 für „weiblich". Die Anbindung an tMap wurde wie in Abb. 18 umgesetzt.

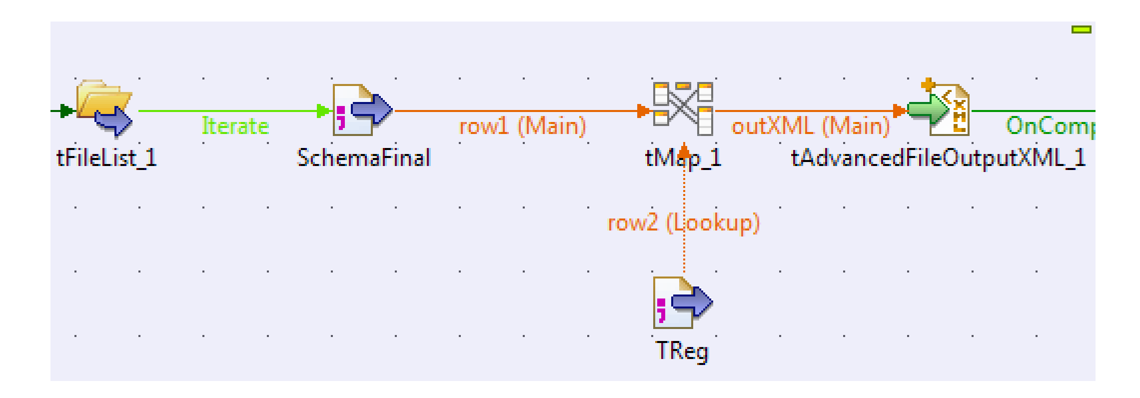


Abb. 18: Anbindung eines zweiten Inputs an eine tMap

Abb. 19 zeigt wie „G“ als Schlüssel auf die andere Tabelle verweist. Wenn der Wert gefunden wird, wird „Geschlecht“ in die andere Tabelle als „G“ übernommen.

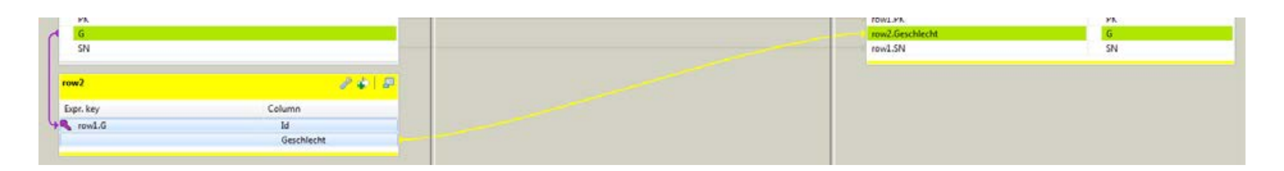


Abb. 19: Transformationsregel Lookup

Regel 2 – Aus mehreren Feldern zusammenlegen

Bei dieser Regel müssen Informationen aus mehreren Zeilen der Ursprungsdatei in einer Zeile zusammengefasst werden. Die kann durch Anwenden der Java-Syntax einfach als String kombiniert werden. Dies ist exemplarisch in Abb. 20 dargestellt. Die Namen N1-N3 sind unter GName zusammengefasst. Die + Zeichen verknüpfen die Strings.



Abb. 20: Zusammenlegen von Zeilen

Regel 3 – Aus einem Feld mehrere erstellen

Diese Regel teilt Informationen, die in einem Feld gespeichert sind auf, und speichert sie in mehreren anderen. Exemplarisch wurde dies mit dem Geburtstag getan (siehe Abb. 21). Die Tage (GDat_Tag), Monate (GDat_Monat) und Jahre (GDat_Jahr) wurden einzeln gespeichert.



Abb. 21: Aufteilen von Datensätzen

8 Prototypentest

In diesem Kapitel werden die zwei erstellten Prototypen nun abschließend getestet. Zunächst werden jedoch die Grundlagen (Standards, Normen, Vorgehensweise) der Tests erläutert.

8.1 Definitionen

Testen: Das Testen ist ein Prozess, der nach dem allgemeinen Standard ISO29119-1 eine Anzahl von verschiedenen Aktivitäten umfasst, mit Hilfe derer die Eigenschaften eines Test-Objekts bewertet werden.[74]

Test-Typen: Grundsätzlich gibt es die vier verschiedenen Testarten[75]:

- Sicherheitstest (*security testing*): Testet den Schutz der Daten und Informationen des Testobjekts gegenüber unautorisierter Systeme und Personen. In diesem Projekt wird kein Sicherheitstest durchgeführt.

- Funktionstest (*functional testing*): Überprüft die vollständige und korrekte Umsetzung der funktionalen Anforderungen an das Testobjekt. Die Testfälle des Funktionstests werden demnach aus den funktionalen Anforderungen abgeleitet. Die zwei in Kapitel 8 genannten Prototypen werden einem Funktionstest unterzogen.

- Gebrauchstauglichkeitstest (*usability testing*):Testet die Bedienbarkeit und Ergonomie des Testobjekts unter Berücksichtigung des Standards EN IDO 9241. Dieser Test wird in diesem Projekt nicht durchgeführt.

- Performance Test (*performance testing*): Testet die Funktionen des Testobjekts unter deren leistungsrelevanten Eigenschaften wie z.B. Antwortzeiten und Durchlaufzeiten. Neben dem Funktionstest, ist dies der zweite Test, in dem die zwei Prototypen getestet werden.

Je nach Test-Typ liegen der Schwerpunkt und die Zielsetzung des Tests auf einem bestimmten Qualitätsmerkmal. Die Qualitätsmerkmale werden nach der DIN 66272 wie folgend definiert[76]:

[74] Vgl. IEEE (2013), S. 12
[75] Vgl. IEEE (2013), S. 2 ff.
[76] Vgl. Franz, K. (2007), S. 14 f.

- Funktionalität (*Functionality*): Die Software weißt die festgelegten und vorausgesetzten Funktionen auf.
- Zuverlässigkeit (*Reliability*): Die Software läuft unter festgelegten Bedingungen und einem festgelegten Zeitraum korrekt und zuverlässig.
- Benutzbarkeit (*Usability*): Die Software besitzt die Fähigkeit für den Benutzer verständlich und bedienbar zu sein.
- Effizienz (*Efficiency*): Die Software liefert mit einem bestimmten Ressourcenverbrauch ein gewisses Leistungsniveau (Performance)
- Änderbarkeit (*Maintainability*): Die Software kann auf sich verändernde Anforderungen abgeändert werden.
- Übertragbarkeit (*Portability*): Die Software kann von einer Umgebung in eine andere organisatorische, Hardware- oder Software-Umgebung übertragen werden.

Testarten: Eine Unterscheidung der verschiedenen Testarten ist mit Blick auf die unterschiedlichen Sichten auf, bzw. in das Testobjekt möglich, den sogenannten Box-Tests. Demnach gibt es drei verschiedenen Boxen[77]:

- White-Box: Im White-Box-Test wird die innere Struktur eines Testobjekts auf Korrektheit und Vollständigkeit geprüft. Hierzu wird der Programmcode analysiert und einzelne Programmabläufe getestet. Daher wird der White-Box-Test auch oft Strukturtest genannt.
- Black-Box: Im Back-Box-Test werden funktionale und nicht-funktionale Anforderungen getestet. Im Unterschied zum White-Box-Test wird hier nicht der Programmcode und innerer Strukturen analysiert, sondern nur die Eingabe und Ausgabe des Testobjekts. D.h. der sieht das Testobjekt nur von außen, also als Blackbox. Der in dieser Arbeit durchgeführte Funktionstest wird nach Verfahren des Black-Box-Tests durchgeführt.
- Grey-Box: Im Grey-Box-Test werden die Sichtweisen des White-Box-Test und des Black-Box-Tests vereint. Somit prüft der Grey-Box-Test die richtige Umsetzung der Anforderungen, den korrekten Ablauf sowie den korrekten Aufbau der inneren Strukturen. Aufgrund des Umfangs, werden Grey-Box-Tests in der Regel nur bei Integrationstest durchgeführt, in denen es darum geht, die Integration und Funktionsfähigkeit eines Systems in ein größeres System zu testen.

[77] Vgl. Franz, K. (2007), S. 27 ff.

8.2 Funktionstest

Im Bereich des Black-Box-Tests gibt es verschiedene Black-Box-Testverfahren. Neben der Äquivalenzklassenbildung, Grenzwertanalyse, Intuitive Testfallermittlung gibt es das Verfahren der Funktionsabdeckung[78]. In anderen Literaturen wird dieser auch oft Funktionaler Systemtest[79] oder Use-Case-Test[80] genannt. Die Idee hierbei ist es, jede funktionale Anforderungen an die Software, bzw. jeden Use-Case zu testen und nachzuweisen, dass diese Funktion vorhanden und ausführbar ist. Es wird demnach das Qualitätsmerkmal „Funktionalität" nach DIN 66272 überprüft. Hierzu werden im ersten Schritt Testfälle und Testszenarien entworfen, mit denen die funktionalen Anforderungen dann im zweiten Schritt systematisch getestet werden können.[81]

8.2.1 Testvorbereitung

Die Testfälle werden auf das Normalverhalten des Testobjekts ausgerichtet, d.h. es werden keine Ausnahmefälle oder Grenzfälle getestet. Basis der Testfälle stellen in der Regel Use-Case-Diagramme oder Aktivitätsdiagramme dar. Anschließend werden die Testfälle in einer testfall-Matrix dargestellt. Diese Matrix zeigt an, welche Testkriterien in einem Testfall gleichzeitig getestet werden können, um die Anzahl der Testfälle zu minimieren.[82] Die Abb. 23 auf Seite 57 zeigt das Ergebnis der der „gekürzten" Testfallmatrix. Dementsprechend sind insgesamt 22 Testfälle nötig, um die Prototypen auf die wichtigsten funktionalen Anforderungen zu testen.

Im Zuge dieser Arbeit wurde zur Veranschaulichung ein Aktivitätsdiagramm (siehe Abb. 22 auf der folgenden Seite) erstellt. Jedoch wurden nicht alle Funktionen daraus in konkrete Testfälle übernommen. Basis der Testfälle waren die Anforderungen aus der Aufgabenstellung der .Versicherung. Diese sind in der Testfallmatrix aufgelistet.

Aktivitätsdiagramm:

[78] Vgl. Frühauf, K./ Ludewig, J./ Sandmayr, H. (2006), zit. nach Hamp, T. (2010), S. 81
[79] Vgl. Franz, K. (2007), S. 102 f.
[80] Vgl. Easterbrook, S. (2012), S. 1
[81] Vgl. Franz, K. (2007), S. 102 f.
[82] Vgl. Rätzmann, M. (2004), S. 144 ff.

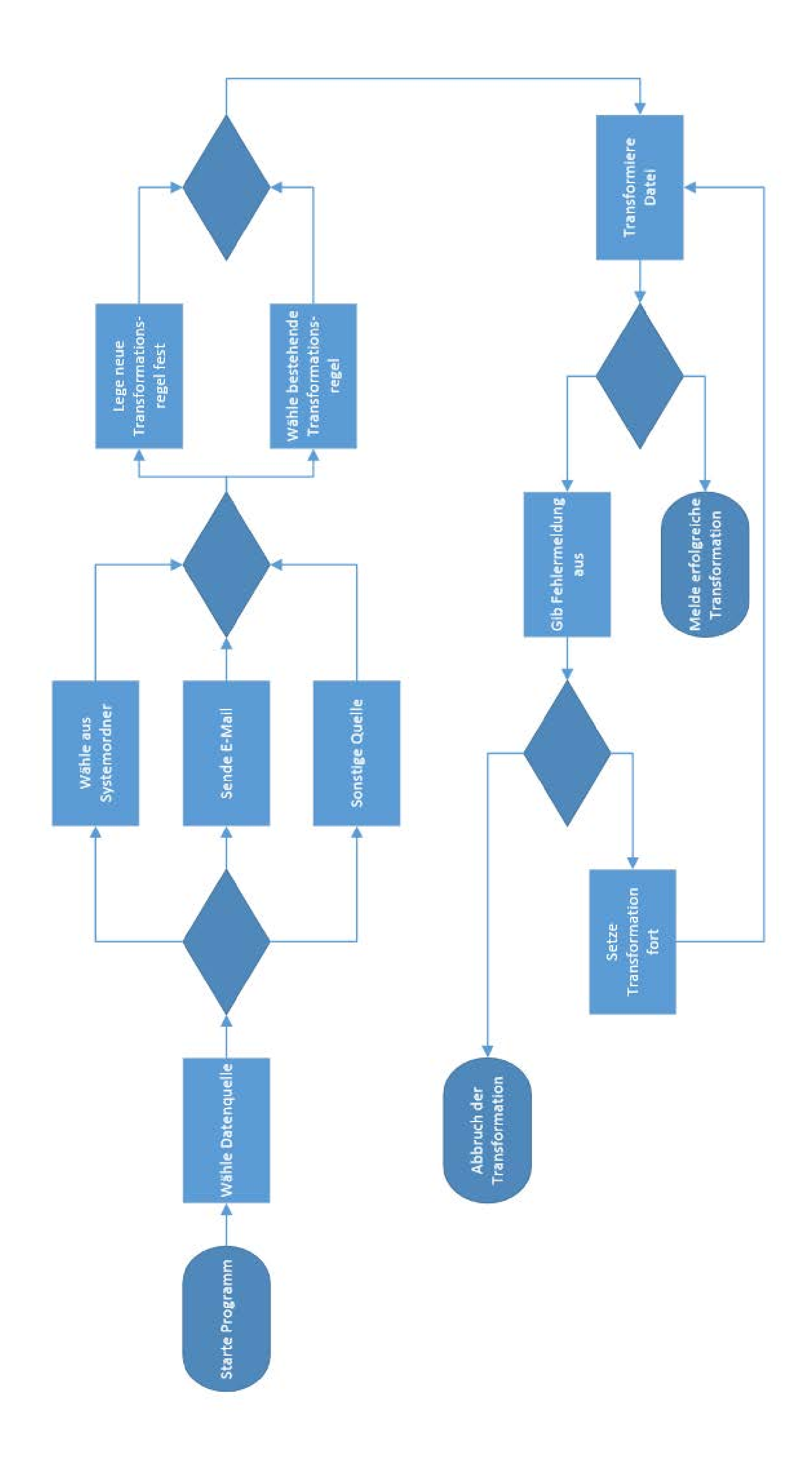


Abb. 22: Aktivitätsdiagramm Funktionstest

<u>Testfallmatrix:</u>

			Testfall																					
Bedingung	Definition	ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	
																									Ergebnis
Transformation	XML -> EDIFACT	T1-1	x																						
	XML -> JSON	T1-2		x																					
	XML -> CSV	T1-3			x																				
	XML -> GDV	T1-4				x																			
	EDIFACT -> XML	T1-5					x																		
	EDIFACT -> JSON	T1-6						x																	
	EDIFACT -> CSV	T1-7							x																
	EDIFACT -> GDV	T1-8								x															
	JSON -> XML	T1-9									x														
	JSON -> EDIFACT	T1-10										x													
	JSON -> CSV	T1-11											x												
	JSON -> GDV	T1-12												x											
	CSV -> XML	T1-13													x										
	CSV -> EDIFACT	T1-14														x									
	CSV -> JSON	T1-15															x								
	CSV -> GDV	T1-16																x							
	GDV -> XML	T1-17																	x						
	GDV -> EDIFACT	T1-18																		x					
	GDV -> JSON	T1-19																			x				
	GDV -> CSV	T1-20																				x			
Regelwerk	Herr/Frau -> 1/0	T2-1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			
	Zusammenlegen Vor-/Nachname	T2-2	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			
	Trennen von Geburtsdatum	T2-3	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			
Attribuieren	Programm lässt sich von außen über Attributes File attributisieren	T3-1																					x		
Monitoring	skalierbare Logging- und Monitoring-Komponenten	T4-1																						x	

Abb. 23: Testfallmatrix

8.2.2 Testdurchführung

Bei der Durchführung des Funktionstests konnten die Testfälle: 1, 5, 6, 7, 8, 10, 14, 18 bei beiden Prototypen nicht durchgeführt werden, da die Transformationen beim Implementieren nicht eingerichtet wurden. Bei diesen Testfällen handelt es sich um Transformationen, bei denen der Datentyp EDIFACT involviert ist. Da der für den Performance-Test genutzte Testdatensatz im Aufbau keinem der verschiedenen EDIFACT-Datensatz-Arten ähnelt, wurden diese Transformationen aus Zeitgründen nicht getestet.

8.2.3 Analyse der Testberichte

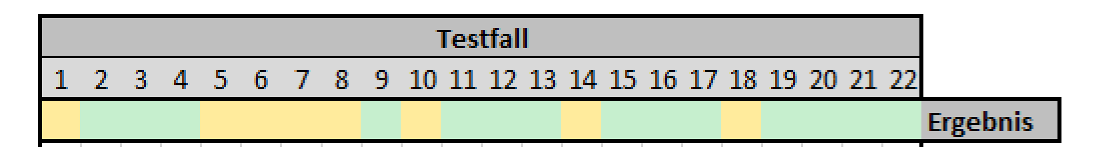


Abb. 24: Ergebnis Funktionstest – Bots

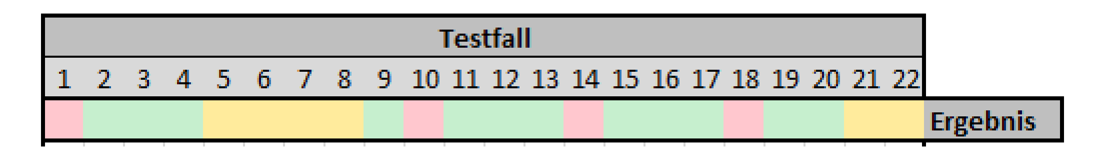


Abb. 25: Ergebnis Funktionstest – Talend

Die Abbildungen Abb. 24 und Abb. 25 zeigen die Ergebnisse des Funktionstests. Grüne Markierungen zeigen einen erfolgreichen Funktionstest. Rote Markierungen bedeuten eine Nichterfüllung der funktionalen Anforderung. Gelbe Markierungen zeigen eine teilweise Erfüllung der Anforderung. Zu sehen ist, dass wie im Abschnitt Testdurchführung schon beschrieben, in beiden Prototypen die Transformationen, bei denen der Datentyp EDIFACT involviert ist, nicht getestet wurden. Laut Herausgeber der Software sind die Transformationen in Bots aber problemlos möglich (siehe Kapitel 6.1.1). In Talents sind jedoch nur Transformationen von EDIFACT in einen anderen Datentyp möglich, jedoch nicht von einem Datentyp nach EDIFACT. Außerdem lässt sich Talend von außen nur eingeschränkt mit einem attributen File attribuieren (Testfall 21). Es sind weitere Konfigurationen im Code nötig. Skalierbare Logging- und Monitoring-Komponenten weißt Talend ebenfalls nur eingeschränkt auf. Der volle Funktionsumfang ist nur in einer kostenpflichtigen Version des Programms erhältlich.

Als Fazit lässt sich aus dem Funktionstest erkennen, dass Bots mehr funktionale Anforderungen erfüllt als Talent. Besonders die Probleme bei der Transformation vom Datentyp EDIFACT können ein entscheidendes Kriterium bei der Wahl der Software darstellen.

8.3 Performance Test

Es gibt verschiedene Vorgehensweisen zum Performance Test. Dieser Performance Test orientiert sich am Vorgehensmodell von Gao, Tsao und Wu. Dieses beschreibt den Performance Test Prozess in acht einzelnen Schritten (siehe Abb. 26Abb. 26) und wird im Folgenden kurz dargestellt. Im ersten Schritt geht es darum, die Leistungsanforderungen genau zu definieren. Dabei hilft es, die SMART-Kriterien (Spezifisch, Messbar, Akzeptiert, Realistisch, Terminiert) hinzuzuziehen. Dieser erste Schritt ist sehr wichtig, um ein aussagekräftiges Testergebnis zu erhalten. Schritte zwei bis fünf können als Testplanung zusammengefasst werden. Hier geht es darum, die Schwerpunkte des Performance Tests zu definieren und anschließend zu priorisieren (Schritt 2). Die Teststrategie (Performance-Maße und Messverfahren), bzw. die Kriterien zu definieren (Schritt 3). Und der Bedarf nach geeignete Test- und

Auswertungs-Tools zu ermitteln, die über entsprechende Daten Sammel - und Monitoring-funktonen verfügen (Schritt 4). Die Ergebnisse dieser Untersuchungen werden in einem Testplan festgehalten (Schritt 5). Der nächste große Schritt ist die Test-Entwicklung. Dafür werden die die Test-Tools eingerichtet, die einzelnen Testfälle (test cases) definiert und Testdaten erstellt (Schritt 6). Der 7. Schritt ist die eigentliche Durchführung des Performance Tests, bei der die Leistungsdaten ermittelt und gesammelt werden. Im letzten Schritt (Schritt 8) werden die gesammelten Daten ausgewertet und grafisch, bzw. strukturiert dargestellt.[83]

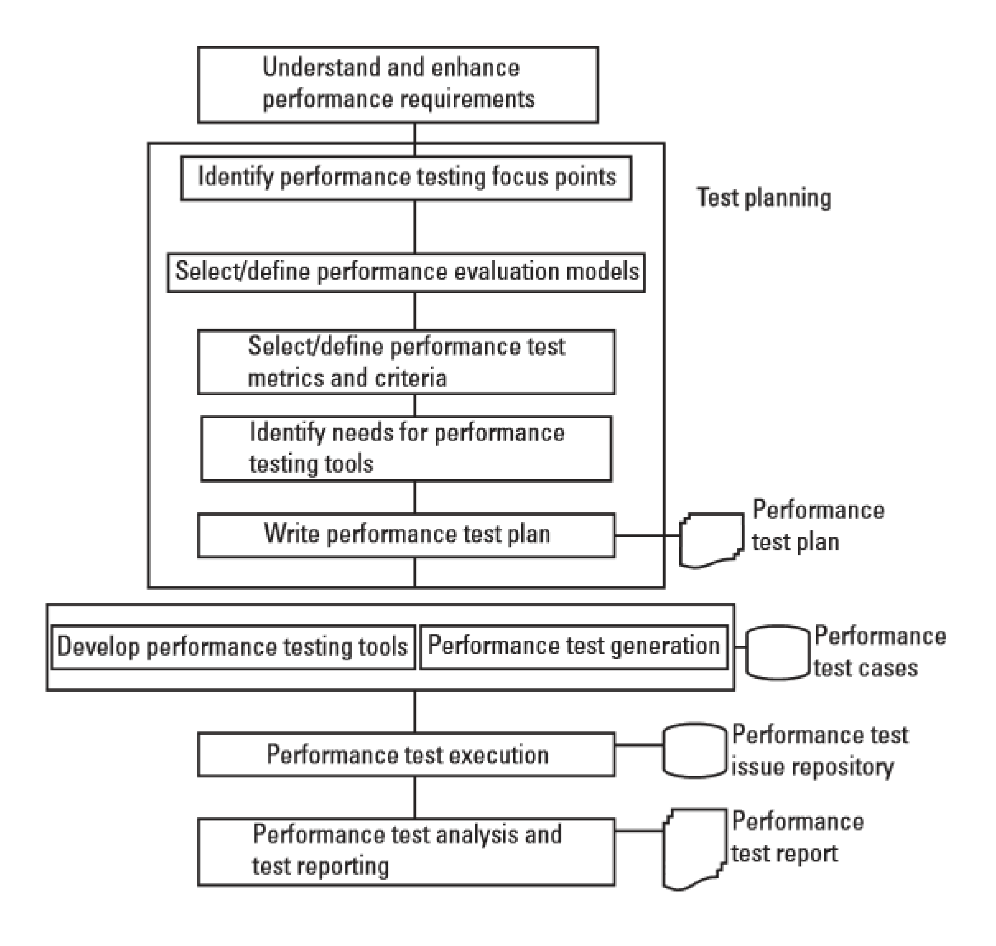


Abb. 26: Vorgehensmodell Performance Test nach Gao, Tsao und Wu[84]

8.3.1 Testplan

- Testziel: Ziel des Performance-Tests ist es, die zwei Prototypen auf die Anforderungen des Partnerunternehmens (.Versicherung) zu testen, um die, mit Blick auf den Datendurchsatz, bessere Software zu ermitteln.
- Testschwerpunkt: Hauptaspekt des Performance-Tests ist das Kriterium Datendurchsatz. D.h. die Anzahl der Dateien, die innerhalb eines bestimmten Zeitrahmens vom

[83] Vgl. Gao, J. Z. / Tsao, H.-S./ Wu, Y. (2003), S. 234 f.
[84] Enthalten in: Gao J. Z./ Tsao, H.-S./ Wu, Y. (2003), S. 234

Datentyp A in den Datentyp B transformiert werden, bzw. die Zeit die zum Transformieren eines bestimmten Datenumfangs (in Gigabyte) benötigt wird. Es wird ebenfalls betrachtet, wie der Datendurchsatz von umfangreichen Transformationsregeln beeinfluss wird.

- Teststrategie: Der Funktionstest aus Kapitel 8.2 hat gezeigt, dass die Transformationen mit dem Datentyp EDIFACT nicht getestet werden können, da hierzu im Aufbau unterschiedliche Testdateien benötigt werden. Beim Verwenden von unterschiedlichen Ausgangsdateien, kann somit kein sinnvoller Vergleich in Bezug auf die Performance der einzelnen Transformationen gezogen werden. Dementsprechend reduziert sich der Testumfang von 240 Testfälle je Prototyp auf 144 Testfälle. Demnach ergibt sich als Ziel des Performance-Tests eine Testabdeckung von 60%. Um frühzeitig Probleme bei hohem Datenvolumen aufzudecken, werden bei beiden Prototypen zunächst die Testfälle mit 1000 bzw. 4000 Dateien getestet.

- Testkriterien: Wie einleitend schon erwähnt wird einzig das Kriterium des Datendurchsatzes bewertet. Berechnet wird der Datendurchsatz in transformierte Datei pro Zeiteinheit ($= \frac{Anzahl\ transformierte\ Dateien}{Zeiteinheit}$). Wobei die Zeit je Datei abhängig von den Faktoren Dateigröße, Datentyp und Komplexität der Transformationsregeln ist.

- Testumgebung: Die Performance Tests werden in der Hardwareumgebungen, die in Tab. 9 dargestellt ist, getestet. Der Testdatenbestand umfasst Testdaten der fünft Datentypen: XML, JSON, CSV, GDV und EDIFACT. Die Anzahl der Testdaten und die Datengröße der einzelnen Testdateien ist in Tab. 11 dargestellt. Der vollständigkeitshalber sind in Tab. 10 die beiden Prototypen nochmals kurz aufgelistet.

	Testumgebung
Betriebssystem	Windows 8.1 (64bit)
CPU	Intel® Core™ i5-2520M CPU (2 Cores, 2.50 GHz)
RAM	4 GB DDR3
HDD	500 GB

Tab. 9: Hardware Testumgebung

	Prototyp A	**Prototyp B**

Name	BOTS	Talend
Lizenz	GPL V3	Apache License, V2.0
Herausgeber	EbbersConsult	Talend Inc.

Tab. 10: Prototypen

Datentyp	**Datengröße je Datei**	**Anzahl der Dateien**	**Datenvolumen (gesamt)**
XML	617 KB	4000	2,47 GB
JSON	629 KB	4000	2,52 GB
CSV	173 KB	4000	692 MB
GDV	253 KB	4000	1,01 GB

Tab. 11: Testdatenbestand

- Bedarf an weiteren Tools: Zunächst wurde überlegt, eine Testsoftware wie den „HP LoadRunner“ oder „Apache JMeter“ zum Messen des Datendurchsatzes zu nutzen. Jedoch wurde festgestellt, dass sowohl Prototyp A als auch Prototyp B in dem jeweiligen Log-Bericht sowohl Start- als auch Endzeitpunkt der Transformation ausgeben. Des Weiteren wurde der Bedarf an einem Daten-Generatoren-Programm untersucht. Da die einzelnen Testdaten jedoch im Aufbau weniger komplex sind und Beispiele im Internet gefunden werden konnten, wurden die Testdateien manuell erstellt. Mit einer einfachen For-Schleife und dem Copy-Befehl lassen sich über die Windows-Konsole Dateien vervielfältigen. In der For-Schleife wird die Anzahl der Dateien angegeben. Die Dateibenennung erfolgt über eine Laufvariable. Hier beispielhaft ein Befehl:

```
for /L %%i IN (2,1,15000) do COPY "dataset1.csv" "dataset%%i.csv"
```

8.3.2 Testvorbereitung

Entwicklung der Testszenarien:

Wie im Testplan schon erwähnt, liegt der Schwerpunkt des Performance-Tests auf der Leistung der Prototypen bei der Transformation der unterschiedlichen Datentypen. Dementsprechend besteht ein Testszenario aus den zwanzig möglichen Kombinationen der Datentypen XML, EDIFACT, JSON, CSV, GDV. Die Transformation von einem Datentyp in denselben unter Berücksichtigung von Transformationsregeln wie z.B. XML nach XML wird nicht betrachtet. Je Transformation gibt es sechs Testfälle, die sich in der Anzahl der zu transformie-

renden Dateien unterscheiden. Diese Einteilung wurde von der .Versicherung wie in Tab. 12 dargestellt festgelegt:

Testfall	Anzahl Testdaten	Gesamtdatenvolumen (Beispiel: GDV)
1	1 Datei	253 KB
2	20 Dateien	5,06 MB
3	100 Dateien	25,3 MB
4	500 Dateien	126,5 MB
5	1000 Dateien	253 MB
6	4000 Dateien	1,01 GB

Tab. 12: Anzahl Testdateien pro Testfall

Insgesamt gibt es zwei Testszenarien. In Testszenario 1 werden alle Transformationen ohne Berücksichtigung von Transformationsregeln durchgeführt. Testszenario 2 beinhaltet ein komplexes Regelwerk. Dieses besteht nach unserer Definition aus drei Transformationsregeln:

1) Substitution: Herr/Frau wird ersetzt mit 0/1
2) Vereinigung: Zusammenlegen der Felder Vorname und Nachname
3) Zerlegung: Trennung des Feldes Geburtsdatum in Tag, Monat, Jahr.

Testdatensatz:

Bei der Auswahl der Testdatei wurde besonders Wert auf den Aufbau der Datei gelegt. Dieser sollte möglichst dem Aufbau der Dateien entsprechen, die in der Versicherungswirtschaft täglich im Einsatz sind. Als Branchenstandard hat sich der GDV-Satz „VU-Vermittler“ etabliert.[85] Die genaue Semantik des Datentyps und der Einsatz in der Versicherungswirtschaft wurde schon in Kapitel 3 beschrieben. Ein entsprechender Datensatz konnte auf der Webseite Gesamtverband der Deutschen Versicherungswirtschaft e. V. gefunden werden. Dieser enthält die Felder: Satzart; VNum; Bündelungskennzeichen; Sparte; Versicherungsschein Nummer; Folgenummer; Geschäftsstelle; Anrede; Name1; Name2; Name3 (Vorname); Titel; Länderkennzeichen; PLZ; Ort; Straße; Postfach; Geburtsdatum; Staatsangehörigkeit; Adresskennzeichen; Aktenzeichen des Sicherungsgläubigers; Zielgruppenschlüssel; Aufsichtsfreier Versicherungsnehmer; Postalisches Kennzeichen; Geschlecht; Satz Nummer.[86]

Die Testdatei enthält 1500 Zeilen. Damit ergibt sich ein Datenvolumen von 253 KB (siehe auch Tab. 11) pro GDV-Testdatei. Um den Anforderung der .Versicherung von 1GB Daten-

[85] Vgl. GDV (2013), S. 9
[86] Vgl. GDV (o. J.)

durchsatz gerecht zu werden, müssen die Prototypen 3953 GDV-Testdaten in einem Transformationsdurchgang bearbeiten.

8.3.3 Testdurchführung

Bei der Testdurchführung wurde frühzeitig erkannt, dass Bots bei großen Datenmengen zum einen sehr lange Zeit benötigt, um diese durchzuführen und zum anderen oftmals nicht alle Dateien transformiert (siehe zum Beispiel Abb. 14 in Anhang 1 – rote Markierung). Dementsprechend wurde regiert und aus Zeitgründen nur die Testfälle mit einer, 20, 100 und 1000 Dateien in Bots durchgeführt. Allein hierfür wurden über 13 Stunden benötigt. Dies entspricht einer Testabdeckung von 50,8%. Bei Talend gab es während der Testdurchführung keine Probleme. Dementsprechend konnten alle 144 Testfälle in einer Gesamtlaufzeit von ca. 2 Stunden 15 Minuten durchgeführt werden.

8.3.4 Analyse der Testberichte

Abb. 27 zeigt exemplarisch einen Ausschnitt aus dem Testprotokoll des Performance-Tests von den Transformationen ohne Regelwerk von XML in JSON, CSV und GDV.

	ohne Regelwerk									
Prototyp	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
	1 Datei		**20 Dateien**		**100 Dateien**		**500 Dateien**		**1000 Dateien**	
Von XML in ...										
XML										
EDIFACT										
JSON	00:00:00	00:00:01	00:00:02	00:00:26	00:00:08	00:02:12	00:00:45		00:01:44	00:22:36
CSV	00:00:00	00:00:01	00:00:02	00:00:23	00:00:08	00:01:46	00:00:44		00:01:30	00:18:14
GDV	00:00:00	00:00:01	00:00:02	00:00:22	00:00:08	00:01:52	00:00:51		00:01:28	00:19:02

Abb. 27: Ergebnis Performance-Test

Schon hier lässt sich deutlich erkennen, dass der Datendurchsatz bei Talend deutlich höher ist als bei Bots. Bei der Transformation von 20 Dateien war Talend beispielsweise ca. 20 Mal schneller als Bots. Der Performance-Unterschied zwischen Bots und Talend wurde in Abb. 28 dargestellt. Da für die Transformation von 500 Dateien in Bots kein Testergebnis vorliegt, wurde ein Zwischenwert von 00:10:00 (hh:mm:ss) gewählt.

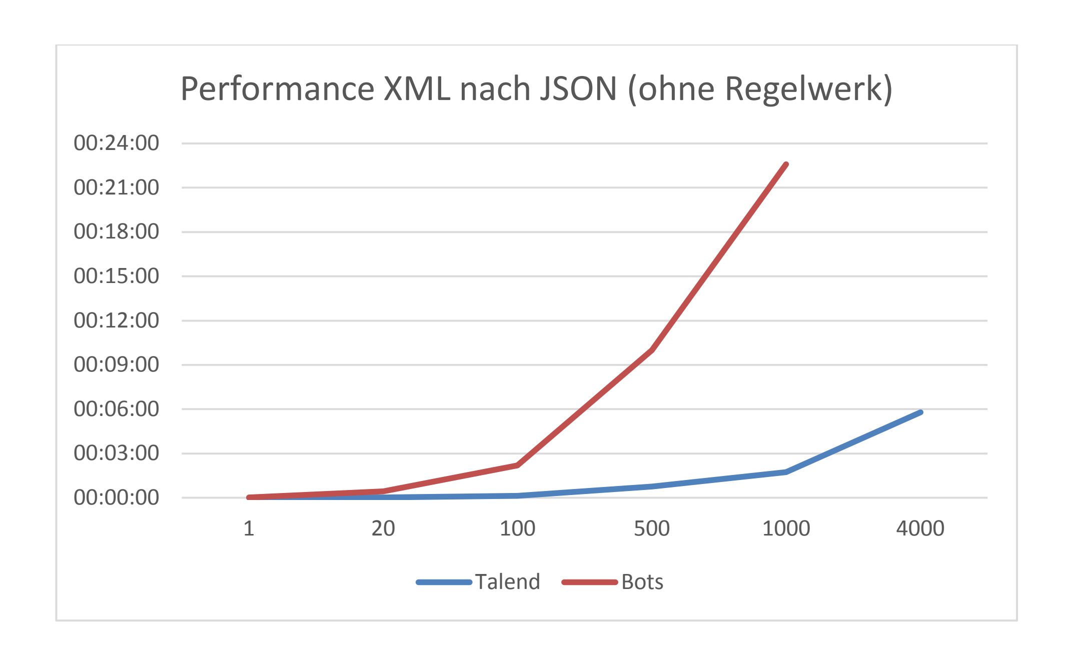


Abb. 28: Diagramm Performance-Test

Ebenfalls bedeutend ist der niedrige Datendurchsatz bei Transformationen von JSON in andere Dateiformate mit dem Tool Talend. Bots hat dagegen eine eher konstante Performance, die wie bereits erwähnt, um ein vielfaches schlechter ist, als die von Talend.

Ein weiteres interessantes Ergebnis der Tests ist die Abnahme des Datensatzes beim Einsatz eines komplexen Regelwerks. Bei Talend nimmt die Performance hier durchschnittlich stärker ab als es bei Bots der Fall ist. So braucht Bots beispielsweise für die Transformation von 1000 Dateien von GDV in CSV ohne spezielles Regelwerk ca. 49-mal so lang wie Talend. Bei derselben Transformation unter Verwendung gewisser Regeln ist Talend nur noch ca. 16-mal schneller.

Es ist zu vermuten, dass Bots trotz seiner langen Durchlaufzeiten eher auf komplexe Regeln ausgelegt ist, als Talend. Da letzteres allerdings trotzdem deutlich bessere Werte aufweist, ist es auch im Hinblick auf diese Kategorie empfehlenswert. Insgesamt ist es auffällig, dass einfach strukturierte Datenformate wie CSV (delimited) und GDV (Fixed-Length-Datensätze) wesentlich schneller transformiert werden können als beispielsweise JSON. Dies könnte an der Größe der Dateien liegen. Letztendlich ist auch zu beachten, dass die Durchlaufzeiten je nach Auslastung des Geräts schwanken können.

Die kompletten Testprotokolle zum Performance-Test sind in Anhang 1 zu finden.

9 Fazit

Durch die stark zugenommene Verbreitung des Internets hat sich der Datenaustausch zwischen Unternehmen gänzlich geändert. Während früher alle Informationen eher innerhalb einer Firma zirkulierten und auf die jeweiligen Anforderungen angepasst waren, gibt es heute eine zunehmende Vernetzung nach außen. Dies führt dazu, dass die Datensätze, mit denen man arbeitet, nicht nur im eigenen System kompatibel sein müssen, sondern auch über die Grenzen des Unternehmens hinaus. Wenn dies nicht möglich ist, müssen die Datensätze schnell und kostengünstig konvertiert werden können. Dies ist die Ausgangssituation, von der aus dieses Projekt gestartet ist. Die .Versicherung war auf der Suche nach einer Open Source Lösung, mit deren Hilfe es möglich ist, genutzte Datentypen zu konvertieren.

Zu Beginn wurde eine umfangreiche Internetrecherche durchgeführt. Deren Ziel war es, ein möglichst umfangreiches Bild über die aktuellen Open Source Produkte im Transformationsbereich zu erhalten. Ergebnis dieser Recherche waren sieben Produkte, die im Wesentlichen den Anforderungen entsprachen. Um diese Auswahl weiter eingrenzen zu können, wurde eine Nutzwertanalyse durchgeführt. Das Resultat war, das drei Produkte den Anforderungen voll entsprachen. Die anderen Produkte wiesen Mängel bei den Datentypen auf oder konnten nicht den Geschwindigkeitsanforderungen bei der Konvertierung entsprechen und wurden daher nicht weiter betrachtet. Um das Ergebnis der Nutzwertanalyse zu kontrollieren, wurden die drei verbleibenden Produkte erneut in einer AHP Analyse bewertet. Dadurch wurde das Ergebnis der Nutzwertanalyse bestätigt. Durch beide Analysen hat sich ein Ranking unter den Tools herausgestellt. Das Tool, das am besten geeignet schien, war Talend. Bots war ebenfalls gut geeignet und befand sich auf Platz zwei. Es wurde entschieden, mit diesen zwei Produkten jeweils einen Prototypen zu erstellen um die vom Anbieter angegebenen Daten zu kontrollieren. Im Verlauf des Testdurchgang stellte sich heraus, dass beide Tools, mit Ausnahme von jeweils einer Schwäche, den Erwartungen hinsichtlich Datentyphandling, Datendurchsatz und Implementierung eines Regelwerk, entsprachen. Mit beiden Tools ist es möglich, komplexe Regelwerke zu definieren und durchzuführen. Wiederum hat Bots Performanceschwächen. Im Schnitt brauchte die gleiche Konvertierung, die in Talend durchgeführt wurde, in Bots vier- bis fünfmal länger. Aber auch Talend zeigte eine Schwäche bei dem Datentyphandling. Alle von der .Versicherung geforderten Datentypen konnten zwar konvertiert werden, EDIFACT ließ sich nur als Ausgangsdatentyp verwenden.

Auch wenn beide Open Source Programme nicht vollständig den Anforderungen entsprechen, sehen die Verfasser einen Vorteil bei Talend. Die Geschwindigkeit der Konvertierung und die Einfachheit, mit der Aufgaben erstellt und durchgeführt werden können, hat überzeugt. Für den Fall, dass EDIFACT als Zieldatentyp vernachlässigt werden kann, ist Talend

Data Integration klar als das beste Open Source Programm zur Datentyp-Konvertierung. Bots stellt ebenfalls eine überzeugende Alternative dar und kann für Datensätze bis 1000 hervorragend eingesetzt werden.

Abschließend fällt bei der objektiven Selbstkritik auf, dass das Vorgehen während der Analysephase generell zu subjektiv gewesen sein könnte. Die Nutzwertanalyse und die AHP müssten in einem größeren Maßstab durchgeführt werden, um absolute Objektivität zu gewährleisten. Durch die starke Spezialisierung der EDIFACT Daten konnten diese während des Testens nicht mit einbezogen werden.

Anhang

Anhang 1: Testprotokolle Performance-Test

Zeitformat [hh:mm:ss]												
	komplexes Regelwerk											
Prototyp	A	B	A	B	A	B	A	B	A	B	A	B
	1 Datei		20 Dateien		100 Dateien		500 Dateien		1000 Dateien		4000 Dateien	
Von XML in ...												
XML												
EDIFACT												
JSON	00:00:00	00:00:01	00:00:02	00:00:26	00:00:08	00:02:57	00:00:50		00:01:41	00:26:04	00:05:29	
CSV	00:00:00	00:00:01	00:00:02	00:00:23	00:00:08	00:02:25	00:00:58		00:01:45	00:20:15	00:05:31	
GDV	00:00:00	00:00:01	00:00:02	00:00:22	00:00:08	00:02:22	00:00:45		00:01:40	00:20:21	00:11:31	
Von EDIFACT in ...												
XML												
EDIFACT												
JSON												
CSV												
GDV												
Von JSON in ...												
XML	00:00:01	00:00:01	00:00:12	00:00:35	00:01:05	00:03:41	00:06:50		00:09:26	00:28:12	00:40:23	
EDIFACT												
JSON												
CSV	00:00:00	00:00:01	00:00:08	00:00:21	00:00:40	00:02:26	00:04:17		00:07:02	00:18:49	00:32:59	
GDV	00:00:01	00:00:01	00:00:07	00:00:22	00:00:40	00:02:23	00:03:16		00:07:06	00:20:04	00:26:31	
Von CSV in ...												
XML	00:00:00	00:00:01	00:00:03	00:00:31	00:00:16	00:03:31	00:01:35		00:04:46	00:27:51	00:12:04	
EDIFACT												
JSON	00:00:00	00:00:01	00:00:00	00:00:28	00:00:05	00:03:02	00:00:17		00:01:02	00:23:51	00:03:37	
CSV												
GDV	00:00:00	00:00:01	00:00:00	00:00:25	00:00:07	00:02:31	00:00:33		00:01:28	00:19:42	00:05:00	
Von GDV in ...												
XML	00:00:00	00:00:01	00:00:02	00:00:31	00:00:08	00:02:53	00:00:35		00:01:14	00:28:57	00:06:42	
EDIFACT												
JSON	00:00:00	00:00:01	00:00:00	00:00:25	00:00:04	00:02:22	00:00:28		00:00:47	00:23:04	00:04:33	
CSV	00:00:00	00:00:01	00:00:00	00:00:19	00:00:04	00:01:49	00:00:21		00:01:18	00:21:16	00:03:01	
GDV												

Tab. 13: Testprotokoll Performancetest – komplexes Regelwerk

Zeitformat [hh:mm:ss]												
	ohne Regelwerk											
Prototyp	A	B	A	B	A	B	A	B	A	B	A	B
	1 Datei		20 Dateien		100 Dateien		500 Dateien		1000 Dateien		4000 Dateien	
Von XML in ...												
XML												
EDIFACT												
JSON	00:00:00	00:00:01	00:00:02	00:00:26	00:00:08	00:02:12	00:00:45		00:01:44	00:22:36	00:05:48	
CSV	00:00:00	00:00:01	00:00:02	00:00:23	00:00:08	00:01:46	00:00:44		00:01:30	00:18:14	00:05:15	
GDV	00:00:00	00:00:01	00:00:02	00:00:22	00:00:08	00:01:52	00:00:51		00:01:28	00:19:02	00:05:16	
Von EDIFACT in ...												
XML												
EDIFACT												
JSON												
CSV												
GDV												
Von JSON in ...												
XML	00:00:02	00:00:02	00:00:15	00:00:29	00:01:12	00:02:33	00:05:37		00:10:48	00:20:09	00:39:22	
EDIFACT												
JSON												
CSV	00:00:01	00:00:01	00:00:09	00:00:22	00:00:41	00:01:50	00:03:17		00:06:34	00:17:36	00:31:45	
GDV	00:00:01	00:00:01	00:00:08	00:00:22	00:00:38	00:01:51	00:03:18		00:06:40	00:18:50	00:27:11	
Von CSV in ...												
XML	00:00:00	00:00:01	00:00:02	00:00:31	00:00:09	00:02:32	00:00:39		00:02:36	00:25:13	00:08:11	
EDIFACT												
JSON	00:00:00	00:00:01	00:00:01	00:00:27	00:00:06	00:02:14	00:00:19		00:00:48	00:22:39	00:05:25	01:44:04
CSV												
GDV	00:00:00	00:00:01	00:00:00	00:00:23	00:00:03	00:01:53	00:00:12		00:01:23	00:19:15	00:02:47	01:23:57
Von GDV in ...												
XML	00:00:00	00:00:01	00:00:02	00:00:28	00:00:09	00:02:32	00:00:39		00:01:14	00:25:16	00:07:59	
EDIFACT												
JSON	00:00:00	00:00:01	00:00:01	00:00:25	00:00:05	00:02:07	00:00:19		00:00:39	00:22:36	00:06:02	
CSV	00:00:00	00:00:01	00:00:00	00:00:18	00:00:03	00:01:32	00:00:11		00:00:20	00:16:22	00:03:00	
GDV												

Tab. 14: Testprotokoll Performancetest – einfaches Regelwerk

Quellenverzeichnisse

Literaturverzeichnis

Balsmeier, P. /Borne, B. (1995): National and International EDI, in: International Journal of Value-Based Management, Nr. 1 vom 1995, S. 53–64.

Banai-Kashani, R. (1989): A New Method for Site Suitability Analysis: The Analytic Hierarchy Process, in: Environmental Management, Nr. 6, S. 685–693.

Bray, T. (2014): RFC 7159: The JavaScript Object Notation (JSON) Data Interchange For mat vom 01.03.14.

Bray, T. u. a. (1998): Extensible markup language (XML), in: World Wide Web Consortium Recommendation REC-xml-19980210 vom 16.08.06.

Crockford, D. (2006): RFC 4627: The application/json Media Type for JavaScript Object Notation (JSON) vom 01.06.06.

Farsi, R. (1999): XML, in: Informatik-Spektrum, Nr. 6 vom 22.10.99, S. 436–438.

Franz, K. (2007): Handbuch zum Testen von Web-Applikationen, Testverfahren, Werkzeuge, Praxistipps, 1. Aufl., Heidelberg: Springer-Verlag.

Frühauf, K./Ludewig, J./Sandmayr, H. (2006): Software-Prüfung, Eine Anleitung zum Test und zur Inspektion, 6. Aufl., Zürich: vdf Verlag.

Gao Jerry Zeyu/Tsao, H.-S. /Wu, Y. (2003): Testing and quality assurance for component-based software, Norwood: Artech house, Inc.

Gesamtverband der Deutschen Versicherungswirtschaft e. V. (2013): Der GDV-Datensatz „VU-Vermittler“, Einsatz, Aufbau, Umsetzung, Betriebswirtschaft und Informationstechnologie, 43, 3. Aufl., Berlin: Ausschuss Betriebswirtschaft und Informationstechnologie.

Hoffmann-Walbeck, T. u. a. (2013): Standards in der Medienproduktion, X.media.press, Berlin, Heidelberg: Springer Vieweg.

IEEE (2013): ISO/IEC/IEEE 29119-1, Software and systems engineering - Software testing, 1. Aufl., Genf: IEEE.

UN/CEFACT Syntax Working Group (JSWG) (1998): Electronic data interchange for administration, commerce and transport (EDIFACT) — Application level syntax rules (Syntax version number: 4) — Part 1, Syntax rules common to all parts, together with syntax service directories for each of the parts vom 01.10.98.

Kühnapfel, J. (2014): Nutzwertanalysen in Marketing und Vertrieb, essentials, 9, Wiesbaden: Springer.

Lamata, M. /Peláez, J. (2003): A New Measure of Consistency for Positive Reciprocal Matrices, in: Elsevier - Computers and Mathematics with Applications, Nr. 12, S. 1839– 1845.

Lehmann, F. (1996): Machine-negotiated, ontology-based EDI (electronic data interchange), Electronic Commerce, Berlin, Heidelberg: Springer.

Rätzmann, M. (2004): Software-Testing und Internationalisierung, 1. Aufl., Bonn: Galileo Press GmbH.

Riedl, R. (2006): Analytischer Hierarchieprozes vs. Nutzwertanalyse: Eine vergleichende Gegenüberstellung zweier multiattributiver Auswahlverfahren am Beispiel Application Service, in: Wirtschaftsinformatik als Schlüssel zum Unternehmenserfolg, S. 99–127.

Saaty, T. (1999): Basic Theory of the Analytic Hierarchy Process: How to make a Decision, in: Revista de la Real Academica de Ciencias Exactas, Físicas y Naturales (Esp), Nr. 4, S. 395–423.

Schmalenbach, C. (2007): Performancemanagement für serviceorientierte JAVA-Anwendungen, Werkzeug- und Methodenunterstützung im Spannungsfeld von Entwicklung und Betrieb, 1. Aufl., Heidelberg: Springer-Verlag.

Shafranovich, Y. (2005): Rfc 4180: Common format and mime type for comma-separated values (csv) files vom 01.10.05.

Siriwardena, P. (2014): JWT, JWS and JWE, in: Advanced API Security, S. 201–220.

Unitt, M. /Jones, I. (1999): EDI—the grand daddy of electronic commerce, in: BT Technology Journal, Nr. 3 vom 03.07.99, S. 17–23.

Verzeichnis der Internetquellen

Ausschuss Betriebswirtschaft und Informationstechnologie Gesamtverband der Deutschen Versicherungswirtschaft e. V. (2013): Der GDV-Datensatz „VU-Vermittler“, http://www.gdv-online.de/vuvm/bestand/Broschuere_gdv-datensatz_vu-vermittler.pdf Abruf: 01.02.2015.

Die Beauftragte der Bundesregierung für Informationstechnik, Bundesministerium des Innern (2012): Rechtliche Aspekte der Nutzung, Verbreitung und Weiterentwicklung von Open-Source-Software, http://www.cio.bund.de/SharedDocs/Publikationen/DE/Archi tekturen-und-Standards/migrationsleitfaden_4_0_rechtliche_aspekte_down load.pdf?__blob=publicationFile Abruf: 03.02.2015.

Easterbrook, S. (2012): Lecture 17: Testing Strategies, http://www.cs.toron-to.edu/~sme/csc302/notes/17-Testing2.pdf Abruf: 27.01.2015.

Ebbers, H. (2012a): Bots Wiki, https://code.google.com/p/bots/wiki/StartInstallDependencies Abruf: 03.02.2015.

Ebbers, H. (2012b): Bots Wiki, https://code.google.com/p/bots/wiki/GrammarsIntroduction Abruf: 03.02.2015.

Ebbers, H. (2012c): Bots Wiki, https://code.google.com/p/bots/wiki/DeploymentEngineOver view Abruf: 03.02.2015.

Ebbers, H. (2014a): About Bots edi translator software, http://bots.source forge.net/en/about_features.shtml Abruf: 29.01.2015.

Ebbers, H. (2014b): bots - Bots open source EDI translator - Google Project Hosting, https://code.google.com/p/bots/ Abruf: 29.01.2015.

Ebbers, H. (2014c): ConfigurationHow - bots - Bots open source EDI translator – Google Project Hosting, https://code.google.com/p/bots/wiki/ConfigurationHow Abruf: 29.01.2015.

Ebbers, H. (2014d): People - bots - Bots open source EDI translator - Google Project Hosting, https://code.google.com/p/bots/people/list Abruf: 29.01.2015.

Ebbers, H. (2014e): StartGetBotsRunning - bots - Bots open source EDI translator - Google Project Hosting, https://code.google.com/p/bots/wiki/StartGetBotsRunning Abruf: 29.01.2015.

Ebbers, H. (2014f): StartIntroduction - bots - Bots open source EDI translator - Google Project Hosting, https://code.google.com/p/bots/wiki/StartIntroduction Abruf: 29.01.2015.

Ebbers, H. (o. J.): EbbersConsult: consultancy for EDI, http://www.ebbersconsult.com/en/in dex.shtml Abruf: 29.01.2015.

Ecosio GmbH (2015): Aufbau einer EDIFACT Datei, http://eco sio.com/de/blog/2014/05/15/Aufbau-einer-EDIFACT-Datei/ Abruf: 15.01.2015.

GEFEG mbH (2014): E-Standards für E-Business - EDIFACT-basiert, XML, Datenmodelle, http://www.gefeg.com/de/standard/standards.htm Abruf: 27.01.2015.

Gesamtverband der Deutschen Versicherungswirtschaft e. V. (2015a): Musterdatei Bestandsdaten, http://www.gdv-online.de/vuvm/musterdatei_bestand/mus ter.html Abruf: 01.02.2015.

Gesamtverband der Deutschen Versicherungswirtschaft e. V. (2005): Beispieldateien, http://www.gdv-online.de/vuvm/musterdatei_bestand/musterdatei_041222.txt Abruf: 01.02.2015.

Gesamtverband der Deutschen Versicherungswirtschaft e. V. (2015b): Beteiligungs-In formations-Satz, http://www.gdv-online.de/vuvm/bestand/rel2013/ds0300.htm Abruf: 20.01.2015.

Häge, M. (o.J.): EDIFACT Musterbestellung, http://www.amf.de/de/downloads/how-to-or der/edifact-musterbestellung.pdf Abruf: 15.01.2015.

Information Sciences Institute (o. J.a): Karma, http://www.isi.edu/integration/karma/# Ab ruf: 02.02.2014.

Information Sciences Institute (o. J.b): Web-Karma, https://github.com/usc-isi-i2/Web-Karma/wiki/Installation Abruf: 02.02.2015.

Information Sciences Institute (o. J.c): Web-Karma, https://github.com/usc-isi-i2/Web-Karma/wiki/Batch-Mode Abruf: 02.02.2015.

mbH, G. (2015): EDIFACT & GEFEG.FX - Infos und Daten zum weltweit verbreiteten Standard der UN, http://www.gefeg.com/de/standard/edifact/edifact.htm Abruf: 20.01.2015.

o. V. (o. J.): (99+) Bots Open Source EDI Translator – Google Groups, https://groups.google.com/forum/#!forum/botsmail Abruf: 29.01.2015.

Talend Germany GmbH (o. J.a): Talend, https://de.talend.com/products/data-integration Abruf: 03.02.2015.

Talend Germany GmbH (o. J.b): Talend, https://help.talend.com/dis
play/HOME/Talend+Open+Studio+for+Data+Integration Abruf: 03.02.2015.



Open Source Security: Sicherheit von Linux auf System z

Schriftliche Ausarbeitung
im Rahmen der Lehrveranstaltung „Integrationsseminar“
für das Kompetenzzentrum Open Source (KOS)

Vorgelegt von

Christopher Choinka, Niko Gerlach, Yann Guerraz,
Pascal Müller, Jan-Lukas Wennemar

am 04.02.2015

Fakultät Wirtschaft
Studiengang Wirtschaftsinformatik
WWI2012E

Inhaltsverzeichnis

Abkürzungsverzeichnis

ACL	Access Control List
BSD	Berkeley Software Distribution
CIFS	Common Internet Filesystem
DAC	Discretionary Access Control
DBM	Database Management
DNS	Domain Name System
DHBW	Duale Hochschule Baden-Württemberg
EAL	Evaluation Assurance Level
ExFS	Extended File System
EXT2	Extended File System 2
EXT3	Extended File System 3
EXT4	Extended File System 4
GDBM	GNU database manager
GNU	GNU's Not Unix, Unix-ähnliches Betriebssystem
GPL	General Public License
GUI	Graphical User Interface
IEFT	Internet Engineering Task Force
IFL	Integrated Facility for Linux
JFS	Journaled File System
KVM	Kernel-based Virtual Machine
LDAP	Lightweight Directory Access Protocol
LDBM	Lightweight DBM backend
LGPL	GNU Lesser General Public License
LXC	Linux Containers
MAC	Mandatory Access Control
MLS	Multi-Level Security
NIS	Network Information Service
NSA	National Security Agency

OSD	Open Source Definition
OSI	Open Source Initiative
OSS	Open Source Software
RBAC	Role Based Access Control
ReiserFS	Reiser File System
SASL	Simple Authentication and Security Layer
SELinux	Security Enhanced Linux
SMB	Server Message Block
TE	Type Enforcement
TLS	Transport Layer Security
VM	Virtual Machine
XFS	Extended File System von Silicon Graphics

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die zunehmende Wichtigkeit von Open-Source Plattformen und Anwendungen wird heutzutage in vielen Bereichen der IT bemerkbar. Ein Grund dafür ist die Kosteneinsparung, die durch den Einsatz kostenlos verfügbarer Applikationen erzielt wird. Ein Beispiel hierfür ist das IT-Umfeld der .Versicherung, auf welches sich das Ergebnis dieser Arbeit bezieht. Das bestehende System der .Versicherung benutzt die z/OS Plattform von IBM, welche über das Sicherheitsprodukt RACF die Zugriffssicherheit regelt. Durch den Wechsel auf eine Open-Source Lösung unter Linux sollen erhebliche Kosteneinsparungen erzielt werden, jedoch soll dasselbe Sicherheitsmaß aufrechterhalten, wenn nicht sogar erhöht werden.

Hohe Sicherheitsstandards sind heutzutage in IT-Umgebungen ein Muss. Besonders wenn vertrauliche Kundeninformationen oder Sozialdaten damit in Verbindung stehen, muss ein höchstes Maß an Sicherheit geboten werden. Zu diesem Zweck gibt es zahlreiche Softwarelösungen, welche unter anderem die Datenintegrität, Sicherheit und Zugriffsrechte verwalten und überwachen. Ziel dieser Arbeit ist dementsprechend, die vorhandenen Softwarelösungen bzw. Vorgehensmodelle zur Absicherung vertraulicher Kundeninformationen und sensibler Daten unter Linux zu untersuchen und anhand eines erarbeiteten Kriterienkatalogs zu bewerten.

Mit einer Marktstudie werden das Einsatzspektrum und die Funktion der jeweiligen Softwareangebote dargestellt, um diese Problematik zu lösen. Dabei wird im ersten Schritt des theoretischen Teils der Arbeit der Begriff Open Source definiert und die Geschichte dazu näher erläutert. Ebenfalls damit verbunden sind die Geschichte zu Linux und die generelle Sicherheit dieses Open-Source Betriebssystems, welche im darauf folgenden Teil beschrieben werden. Die Einbindung von Linux auf dem System z von IBM, dem sogenannten z/Linux, wird am Ende des Theorieteils untersucht, um ein besseres Verständnis über das System zu bieten.

Der Kriterienkatalog, welcher benutzt wird, um die Softwarelösungen zu bewerten, wird im darauf folgenden Teil vorgestellt. Die Kriterien werden durch die Beleuchtung der Ist-Situation bei der .Versicherung abgeleitet. Im nächsten Schritt wird das generelle Bewertungsverfahren präsentiert, mit dem die einzelnen Sicherheitsanwendungen ausgewertet werden. Im Anschluss daran, werden die einzelnen Sicherheitslösungen vorgestellt und bewertet. Dabei wurden die aktuellsten verfügbaren Anwendungen gewählt. Diese sind Access Control Lists in Extended File Systems (ACL), Samba, Linux Container (LXC), Lightweight Directory Access Protocol (LDAP) und Security Enhanced Linux (SELinux).

Im finalen Abschnitt werden die ausgearbeiteten Ergebnisse konsolidiert und kritisch analysiert. Hierbei steht im Vordergrund eine Antwort auf die Frage zu finden, ob es möglich ist, unter Linux ein sicheres Dateisystem sowie eine präzise Zugriffsrechtevergabe aufzusetzen.

2 Open Source

Im Folgenden erfolgt die Darstellung der Geschichte, Definition und Abgrenzung der Begrifflichkeit Open Source.

2.1 Geschichte

Anfang des Jahres 1998 befand sich das Unternehmen Netscape in einer wirtschaftlich schwierigen Lage. Es hatte den Konkurrenzkampf mit Microsofts Internet Explorer verloren. Netscapes damaliges Hauptprodukt Netscape-Navigator war gescheitert.[1] Das Unternehmen entschied sich daher den Quellcode ihres Produktes öffentlich verfügbar zu machen. Schnell interessierten sich Gruppen freier Software Programmierer und GNU (GNU's not Unix) für das Produkt. Das erste Open Source Projekt mit weltweiter Beachtung war geboren.

Zwar gab es zuvor bereits Open Source Projekte wie Unix 1969 oder GNU 1984, jedoch nicht mit dem Bekanntheitsgrad der Netscape Pleite.

Zeitgleich kristallisierten sich 1998 die Hauptbestandteile der Open Source Definition heraus. Abgeleitet wurden sie von der Linux-Distribution Debian und dem Debian Social Contract, die im Vorjahr entstanden waren. Im gleichen Jahr wurde die Open Source Initiative (OSI) gegründet. Dieser Initiative wurde nach der Gründung das Certification Mark (CT) für den Namen Open Source von der Organisation Software in the Public Interest sowie die Verantwortung der Open Source Definition und Lizenzen übertragen. Mit diesen Ereignissen war das Fundament von Open Source Software begründet sowie rechtlich abgesichert.

2.2 Definition

Bereits in dem vorrangegangen Kapitel ist der Ursprung der Definition von Open Source dargelegt worden. Die Open Source Definition (OSD) der Open Source Initiative (OSI) kann auf der Website gefunden werden und beginnt wie folgt:

„Open source doesn't just mean access to the source code. […]“ [2]

Der erste Satz dieser Definition kann bei dem Leser die Frage auslösen, warum die Definition in Form des Ausschlussprinzips stattfindet und beschreibt was Open Source nicht ist. Dieser Satz ist jedoch äußerst wichtig für das Verständnis der Grundidee von Open Source.

[1] Vgl. Heise (1998)
[2] Vgl. Open Source Initiative (o.J.a)

Hinter dem Begriff verbirgt sich nicht nur der öffentliche Zugang zu dem Quellcode, sondern weitere neun Kriterien müssen erfüllt sein, damit ein Produkt mit der geschützten Marke O-pen Source bezeichnet werden darf. Auch diese Kriterien sind auf der Website von o-pensource.org hinterlegt. Sie sind der Tabelle „Kriterien zur Bestimmung eines Open Source Produktes“ (s. Anhang Nr. 1) zu entnehmen.[3]

2.3 Abgrenzung Free Software vs. Open Source

Die Begrifflichkeiten Free Software und Open Source werden häufig synonym verwendet, daher soll im Folgenden eine kurze Gegenüberstellung der Begriffe erfolgen.

Der größte Unterschied der beiden Begriffe ist die Prägung durch zwei unterschiedliche Organisationen. Der Begriff Free Software ist von der Free Software Foundation geprägt worden, der Begriff Open Source von der Open Source Initiative. Open Source betont den Entwicklungsprozess und die Quelloffenheit. Die Free Software basiert auf den vier Grundsätzen der Freiheit ein Programm nachvollziehen, verändern, verteilen und beliebig nutzen zu können.[4] Dies sind die Hauptunterscheidungspunkte der zwei Bewegungen.

[3] Vgl. Open Source Initiative (o.J.a)
[4] Vgl. Lang, M. (2012)

3 Linux

Linux ist ein Mehrbenutzer-Betriebssystem, das unter der Open Source-Lizenz GPL vertrieben wird. Es basiert wesentlich auf GNU Software und dem Linux-Kernel, die vollständig frei erhältlich sind. Linux ist in vielen verschiedenen Distributionen für verschiedene Endanwender wie Privatpersonen oder Unternehmen verfügbar. Es kann auf allen möglichen Geräten wie Servern, Desktop-PCs, Laptops und Smartphones installiert und dort genutzt werden.[5] In diesem Kapitel werden kurz die Entwicklungsgeschichte und die Hauptfunktionen des Kernels beleuchtet. Außerdem werden die Extended File Systems unter Linux vorgestellt.

3.1 Geschichte

Start der Geschichte von Linux ist die Gründung des GNU-Projekts im Jahr 1982 durch Richard Stallman. Im Zuge dieses Projekts sollen verschiedene Open-Source Werkzeuge für das Betriebssystem Unix entwickelt werden, da viele Werkzeuge durch die Lizensierung von Software nicht frei verfügbar sind. 1989 wird eine neue Lizenz durch die GNU-Verantwortlichen eingeführt: die GPL (General Public License). Mit dieser Lizenz ist es jedem erlaubt, die Software zu nutzen, zu ändern und zu verbreiten. Außerdem stellt sie sicher, dass jede Erweiterung des Codes ebenfalls unter der GPL vertrieben werden und somit weiterhin frei verfügbar sein muss.[6]

Im Jahr 1991 kündigt Linus Torvalds die Entwicklung eines Open Source Kernels an, der unter dem Namen Linux in der Version 0.01 zur Verfügung gestellt wird. Es werden Erweiterungen, wie Hardwaretreiber oder Filesystemverwaltung, durch andere Entwickler hinzugefügt. Mit der Lauffähigkeit des GNU-C-Compilers auf dem Kernel kann die gesamte Software des GNU-Projektes unter Linux genutzt werden und macht Linux so zu einem vollständigen Betriebssystem.[7]

In den folgenden Jahren werden weitere Entwicklungen am System vorgenommen, sodass 1994 die ersten Distributionen (Debian, Red Hat, Slackware, SUSE) von Linux auf den Markt kommen. Sie verpacken das Betriebssystem in leicht zu installierende Dateien, sodass Linux einer großen Anzahl an neuen Anwendern zugänglich gemacht wird und für Unternehmen eine Alternative zum bisherigen Windows darstellt. Heute werden diese Distributionen sowohl von freiwilligen Entwicklern als auch von Firmen-Entwicklern erweitert. Der kommerzielle Erfolg von Distributionen wie Red Hat wird durch den zusätzlichen Verkauf von Support-

[5] Vgl. Kofler, M. (2008), S. 25
[6] Vgl. Kofler, M. (2008), S. 40
[7] Vgl. Kofler, M. (2008), S. 41

paketen für Endanwender sichergestellt.[8] Durch den steigenden Kostendruck für Unternehmen rückt Linux immer weiter in den Mittelpunkt, da dort keine Lizenzkosten anfallen.[9] Ein zentraler Aspekt von Linux ist der Kernel, der im Nachfolgenden beleuchtet wird.

3.2 Der Kernel

Der Kernel übernimmt auch bei Linux die typischen Funktionen in einem Betriebssystem. Die Hauptfunktionen werden in Abbildung 1 gezeigt.

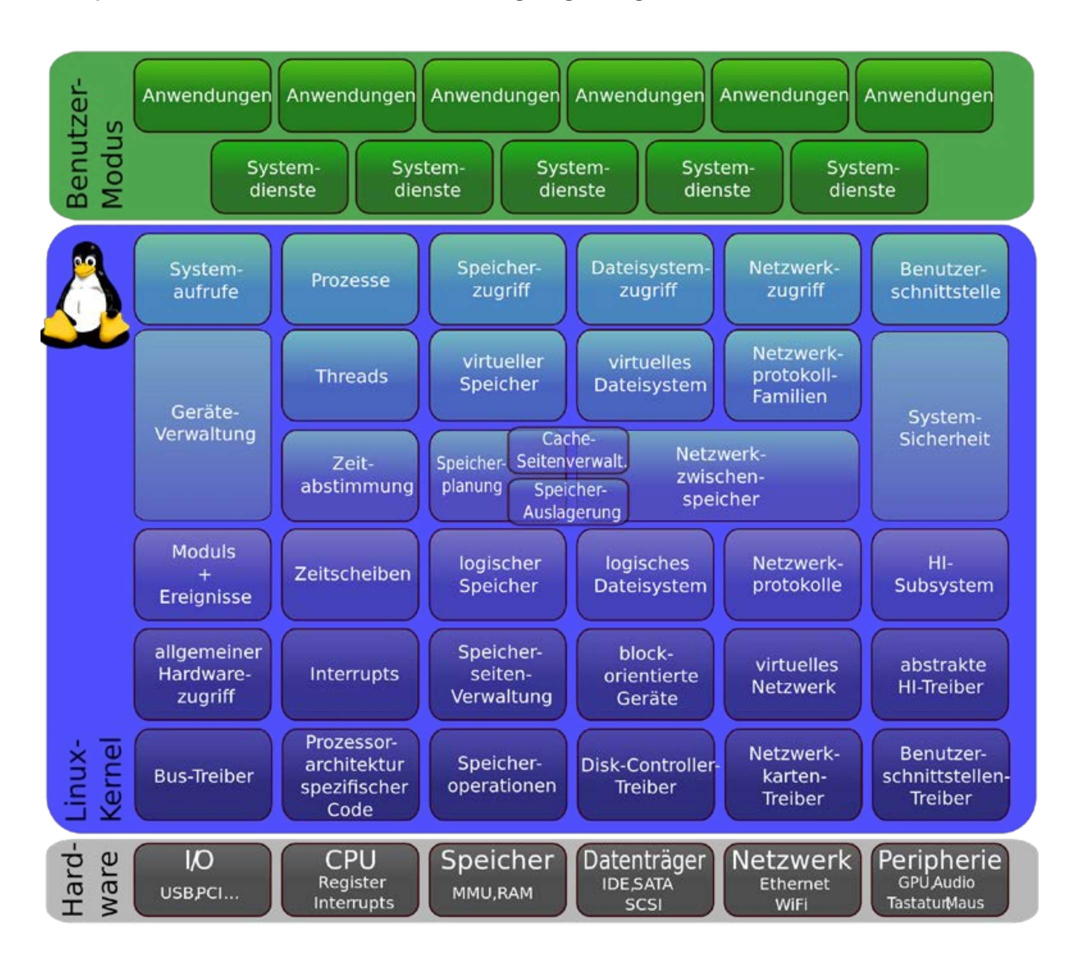


Abb. 1: Funktionen des Linux-Kernels[10]

Die Funktionen des Kernels umfassen generell die gesamte Kommunikation zwischen der Hardware (Speicherverwaltung, Hardwarezugriff, Geräteverwaltung etc.) und der Software bzw. dem Betriebssystem (Threads, Dateisystemzugriff, Netzwerkprotokolle, etc.). Da Linux ein Open Source Produkt ist, arbeiten viele Programmierer weltweit daran, den Funktionsum-

[8] Vgl. Kofler, M. (2008), S. 41
[9] Vgl. Tobler, M. (2001), S. 8
[10] Enthalten in: http://upload.wikimedia.org/wikipedia/commons/thumb/4/46/Linux_Kernel_Struktur.svg/2000px-Linux_Kernel_Struktur.svg.png

fang und auch die Sicherheit des Betriebssystems und im speziellen des Kernels zu verbessern. Dadurch umfasst der Linux-Kernel eine Vielzahl von Besonderheiten:

- Der Kernel unterstützt eine sehr große Zahl an Hardwarekomponenten, sodass fast alle Hardwarekonfigurationen unter Linux verwendet werden können.[11]
- Multitasking, Multiuser-Betrieb sowie Paging (virtuelle Speicherverwaltung) und Shared Libraries (zentral verwaltete Bibliotheken) können vom Kernel umgesetzt werden.[12]
- Außerdem werden verschiedene Dateisysteme (ext2, ext3, JFS usw.) unterstützt.[13]

Durch diesen großen Funktionsumfang und die hohe Kompatibilität kann Linux in fast jedem Umfeld eingesetzt werden und stellt gegenüber den lizensierten Betriebssystemen anderer Anbieter eine gleichwertige Alternative dar.

3.3 Extended File Systems unter Linux

Ein Extended File System oder auch Journaling-Filesystem ist ein Dateiverzeichnis, das alle Veränderungen als Transaktionen im System verbucht und abarbeitet. Bevor die Daten geschrieben werden, werden alle Schritte in ein Journal (reservierten Speicherbereich) geschrieben und gespeichert. Fällt nun das System aus, kann die Transaktion entweder abgeschlossen oder noch nicht abgeschlossen sein. In beiden Fällen bleiben die Daten im System konsistent, da bei nicht abgeschlossener Transaktion der gesamte Vorgang abgebrochen und alles zurückgesetzt wird. Besonders bei großen Datenmengen und Speichermedien ist dies von Vorteil und erspart viel Zeit beim Zurücksetzen.[14]

Beim Metadaten-Journaling werden die Informationen zu einer Datei gespeichert, wie Speicherplatz und Dateigröße, etc. Darüber hinaus können die Zugriffsrechte in einem Extended File System über Access Control Lists (ACLs) vergeben und kontrolliert werden.

[11] Vgl. Kofler, M. (2008), S. 26
[12] Vgl. Tobler, M. (2001), S. 5
[13] Vgl. Kofler, M. (2008), S. 26
[14] Vgl. LinuxWiki (2013)

4 Security unter Linux

Heutzutage sind mehr Computersysteme an das Internet angeschlossen, als jemals zuvor in unserer Geschichte und die Zahl dieser Systeme wächst stetig. Allerdings ist das Internet keineswegs so sicher wie manchmal angenommen wird. Angriffe auf Großunternehmen wie zuletzt auf die Spieleplattformen von Sony und Microsoft belegen dies und lassen nur den Schluss zu, dass es kein komplett sicheres Computersystem gibt, welches vor jeglicher Art von Komprimittierung geschützt ist.[15]

Zu beachten ist, wenn es um die Sicherheit von Computersystemen geht, dass IT-Security sowohl deutlich mehr als Antivirenprogramme und Firewalls umfasst, als auch bereits deutlich früher einzuordnen ist. Bereits auf Betriebssystemebene gibt es Unterschiede in der Sicherheitsqualität, beispielhaft anhand von Linux und Windows gezeigt. Windows-Systeme sind hierbei deutlich häufiger Ziel von Hackerangriffen als Systeme die unter Linux laufen, was nur zum Teil an der deutlich größeren Verbreitung liegt. Naturgemäß werden solche Systeme für einen Angriff ausgewählt, die eine möglichst hohe Erfolgswahrscheinlichkeit versprechen. Neben der weiten Verbreitung von Windows ist hierfür vor Allem die deutlich größere Anzahl an potentiellen Sicherheitslücken verantwortlich, worauf auch die deutlich höhere Patch-Frequenz von Windows hindeutet.[16] Dies lässt allerdings nicht zwingend den Schluss zu, dass Linux ein von sich aus sicheres Betriebssystem darstellt. Bei Linux geht im Gegensatz zu Windows eine größere Gefahr von den Usern selbst als von externen Angreifern aus, welche es durch unterschiedliche Sicherheitsmechanismen zu kontrollieren gilt.[17]

Die größere von Usern bzw. geringere von Hackern ausgehende Gefahr ist im Aufbau von Linux und seiner Art mit Dateien umzugehen begründet, die es Hackern z.B. deutlich schwieriger macht mit Social Engineering in fremde Netzwerke einzudringen, da hierbei deutlich mehr eigene, bewusste Handlungen vom User selbst nötig sind als dies bei Windows der Fall ist. Die daher notwendige User-Sicherheit wird bei Linux durch Authentisierung und Zugriffskontrolle erreicht. Authentisierung kontrolliert dabei, ob ein User, der eine Systemanfrage schickt, wirklich der ist, für den er sich ausgibt, während die Zugriffskontrolle die verschiedenen Rechte eines Users auf unterschiedliche Dateien und Informationen regelt – bei Linux sind dies lesen, schreiben und ausführen bzw. read, write and execute. Ein wichtiger Aspekt, um Gefahren durch neue User auszuschließen ist es, standardmäßig nur die kleinsten notwendigen Rechte zu vergeben, also ausschließlich Leserechte, sofern andere Rechte

[15] Vgl. Linuxtopia (o.J.a)
[16] Vgl. Linuxtopia (o.J.b)
[17] Vgl. Prakasha, S. (o.J.)

nicht ausdrücklich notwendig sind. Dadurch ist es Benutzern, die nur Leserechte besitzen nicht möglich Schadsoftware auszuführen.[18]

Eine weitere äußerst wichtige Facette der User-Security ist die Root-Security, also der Umgang mit Root-Rechten. Da der Root-Account Kontrolle über das gesamte System hat, ist dieser besonders kritisch und im Umgang mit ihm Vorsicht geboten, da sich für unerfahrene Nutzer hier ein großer Nachteil von Linux zeigt: Linux hat keine Kontrollabfragen. Aufgrund dieser Gegebenheit ist es unter dem Root-Account möglich, mit einem einzelnen Befehl das gesamte System zu löschen, ohne dass dies im ersten Moment bemerkt wird. Aus diesem Grund ist es wichtig, den Root-Account nur für spezifische Aufgaben zu verwenden und danach sofort zu einem normalen User zurückzukehren. Der Root-Account ist demnach aufgrund seiner ganzheitlichen Rechte der User mit dem größten Schadenspotential.[19] Auch wenn diese Ausgliederung der Root-Rechte auf einen speziellen Account sich, wie oben geschildert, zuerst nachteilig anhört, stellt sie doch einen weiteren Vorteil von Linux gegenüber Windows in Bezug auf die Systemsicherheit dar. Dadurch, dass ein normaler Linux-User standardmäßig keine allumfassenden Rechte besitzt, ist es z.B. externer Schadsoftware, die diesen Account infiziert hat nicht möglich Kontrolle über das gesamte System zu übernehmen. Windows-User hingegen haben standardmäßig Administratorrechte und können somit auf das gesamte System zugreifen, wie es hier auch potentieller Schadsoftware möglich wäre. Des Weiteren besteht durch Administratorrechte bei Windows für jeden Nutzer die Möglichkeit unabsichtlich großen Schaden am System anzurichten, wodurch sie nur durch mögliche Kontrollfragen abgehalten werden können.[20]

Neben den oben beschriebenen Authentisierungs- und Zugriffskontrollmaßnahmen, welche für diese Seminararbeit maßgeblich sind, gibt es natürlich auch unter Linux noch weitere Sicherheitsmaßnahmen. Dies sind zum einen physische, Passwort- bzw. Verschlüsselungs-Security und Kernel-Security und zum anderen die Netzwerk-Security. Unter der physischen Sicherheit kann u.a. grundlegend verstanden werden, wer physischen Zugriff auf die Systeme hat z.B. durch Separationsschleusen am Gebäudeeingang und Schlössern an den IT-Systemen, welche sie physisch an einen Ort binden. Auch kann darunter das Aufsetzen eines Bootpassworts verstanden werden, was auch in den Bereich der Passwort- oder Verschlüsselungssicherheit reicht. Hierbei kann der unerlaubte Zugriff auf Daten und Informationen z.B. durch Vorgaben der Passwortstruktur und das Aufsetzen einer Festplattenverschlüsselung erschwert werden. Kernel-Security meint die Art, in welcher der Kernel konfiguriert wird um Angreifern den Zugriff auf diesen zu erschweren. Da der Kernel den Informationsfluss des Computers kontrolliert ist es besonders wichtig, diesen zu schützen, z.B. indem

[18] Vgl. Prakasha, S. (o.J.)
[19] Vgl. ebenda
[20] Vgl. Noyes, K. (2010)

dieser mit Software wie SELinux erweitert wird. Die letzte Kategorie „Netzwerk-Sicherheit" beschreibt alle Maßnahmen, die das Netzwerk nach außen hin absichern. Dies können einfache Firewalls sein, aber auch Virtual Private Network Connection zum Zugriff auf das Firmennetzwerk und der Einsatz von Sniffingtools, um den Netzwerktraffic auf bestimmte Befehle wie „password" und „login" hin abzuhören und daraufhin zu überprüfen.[21]

[21] Vgl. Fenzi, K. / Wreski, D. (2004)

5 z/Linux – Funktionalitäten und Sicherheit

Im folgenden Abschnitt wird z/Linux vorgestellt, welches die Vorteile des Open Source Betriebssystems Linux und die des System z von IBM vereint. Die Einbindung von Linux in das z System von IBM bringt viele Funktionalitäten und Sicherheitsstandards, die optimal in IT-Infrastrukturen eingesetzt werden können und diese verbessern. z/Linux bietet hohe Leistungskapazitäten auf einem einzelnen physikalischen Server und fördert die Softwareentwicklung durch moderne Virtualisierungsmethoden. Durch die Erfüllung von Dienstleistungen über eine Private Cloud wird ein höherer Grad der Flexibilität innerhalb der IT-Infrastruktur erreicht. Die Verarbeitung von analytischen Informationen und Daten erfolgt hierbei in Echtzeit. Der wichtigste Punkt ist das hohe Maß an Sicherheit und Zuverlässigkeit, welches durch z/Linux geboten wird.[22]

Durch den Einsatz von z/Linux können erhebliche Kosteneinsparungen erzielt werden. Hunderte virtuelle Linux-Server können parallel und voneinander isoliert über z/Linux betrieben werden. Weitere Kosten können im Bereich der Softwarelizensierung, Wartung/Management, Sicherheit und Energieverbrauch/Platzbedarf eingespart werden. Der Preis von Linux-Software wird im Normalfall pro IFL-Mainframe (Integrated Facility for Linux) berechnet. Durch die parallele Ausführung mehrerer virtueller Linux Server auf diesem System werden somit Kosten eingespart. Durch die Konsolidierung mehrerer Linux-Server auf einem physikalischen Server ist die Anzahl der zu wartenden Hardwarekomponenten geringer. Dies resultiert in einfacheren IT-Infrastrukturen, deren Verwaltungskosten geringer sind. In Kombination mit z/VM (Virtual Machine) bietet z/Linux Sicherheitsmechanismen gegen Transferspitzen und Systemabstürze, was eine erhöhte Geschäftskontinuität sicherstellt. In Verbindung mit z-Servern, können außerdem Gefahren für die Integrität des Systems frühzeitig erkannt und behoben werden. Der Platzbedarf für physikalische Server wird durch den Einsatz von z/Linux drastisch reduziert. Ebenfalls reduziert werden dadurch die Kosten für die Klimatisierung der Räume und den Energieverbrauch.[23]

Diese Einsparungen können anhand der bereits durchgeführten Optimierungen der Shelter Mutual Insurance Company und EFiS EDI Finance Service AG verdeutlicht werden.

Die Shelter Mutual Insurance Company ist eine US-Amerikanische Versicherungsfirma (Missouri) mit ca. 3.700 Angestellten. Steigende Kosten und die Komplexität der Serverlandschaft waren der Hauptgrund für den Einsatz einer z/Linux Lösung. Hierbei war das Ziel die Kosteneinsparung und Vereinfachung der IT-Infrastruktur. Durch die Installation von virtuellen Linux Umgebungen auf IBM zEnterprise 114-Servern konnten tausende US-Dollar pro

[22] Vgl. IBM (2013a)
[23] Vgl. ebenda

Jahr in Lizenzkosten gespart werden. Außerdem wurde die Flexibilität erhöht und ein zentralisiertes Verwaltungssystem aufgebaut.[24]

Die EFiS EDI Finance Service AG ist ein Mitglied der Paymentgroup, welche 73 Mitarbeiter hat und jährlich €10 Millionen Umsatz erzeugt. Das Unternehmen braucht kraftvolle Computer-Systeme, die gleichzeitig stabil genug sind, um dauerhaft im Einsatz zu sein und Technologie besitzen, um hohe Auslastungen zu vertragen. Hierbei steht die Erhöhung der Effizienz und Flexibilität der Computer-Systeme im Vordergrund. Als Lösung wurde IBM z Solution Edition for Enterprise Linux auf 114 IBM zEnterprise Servern aufgespielt. Die Prozessgeschwindigkeit wurde durch das Ersetzen veralteter Hardware verdoppelt und mehr als 30% Kosteneinsparungen beim Energieverbrauch und Softwarelizensierungen konnten festgestellt werden.[25]

Das wichtigste Thema wenn es um z/Linux geht, ist die Sicherheit. Durch die Kombination aus Linux und den hardware-basierten Sicherheitsfunktionalitäten von System z wird das Ausführungsumfeld noch sicherer gemacht. Um ein hohes Maß an Sicherheit zu bieten, wird in System z fast jede Ebene, d.h. der Prozessor, das Betriebssystem und die Anwendungsebene abgesichert. Die Sicherheitsmaßnahmen um vertrauliche Daten und Informationen zu verarbeiten wurden von der International Standards Organization mit dem EAL5-Level zertifiziert, welches die dritthöchste Sicherheitsstufe ist. Des Weiteren werden komplexe Verschlüsselungs-Algorithmen verwendet, um den externen Datentransfer abzusichern. [26]

Eine Vielzahl verschiedener Sicherheitstools wird sowohl von IBM, als auch von Open-Source Projekten und anderen Softwarehändlern für z/Linux angeboten. Beispielsweise stellt IBM Verschlüsselungssoftware (IBM CEX3C) zur Verfügung[27], wodurch der Datentransfer abgesichert werden kann. Zusammen mit zusätzlichen Firewalls[28], werden Webserver und Datenbanken optimal geschützt.[29]

[24] Vgl. IBM (2012a)
[25] Vgl. IBM (2012b)
[26] Vgl. IBM (2012c)
[27] Vgl. ebenda
[28] Vgl. z/Journal (2008), S.6
[29] Vgl. IBM (2013b)

6 Bewertungskatalog und Bewertungsverfahren

In diesem Kapitel sollen die gewählten Bewertungskriterien sowie das gewählte Bewertungsverfahren für die vorliegende Ausarbeitung vorgestellt werden. Zunächst werden die einzelnen Bewertungskriterien vorgestellt, in Gruppen zusammengefasst und anschließend mit der Bewertungsmethode kombiniert. Dabei ist zu beachten, dass im Umfang dieser Arbeit die Analyse der einzelnen Thematiken von einem eigenständigen Teammitglied analysiert wird. Die Bewertungen entstehen anschließend als Gruppendiskussion um eine Verhältnismäßigkeit zwischen den unterschiedlichen Lösungen zu gewähren. Dies hat zur Folge, dass die Bewertung auf der Analyse des Einzelnen und den daraus gewonnen Eindrücken basiert. Um die Objektivität an dieser Stelle zu gewährleisten soll im Verlauf dieses Kapitels eine Möglichkeit vorgestellt werden die Subjektivität in der Bewertung zu vermeiden.

Im Rahmen einer Ausarbeitung des Projektteams sind insgesamt 15 bewertbare Kriterien festgelegt worden. Dabei wurde eine Ausarbeitung der kooperativen Forschung an der dualen Hochschule Baden-Württemberg aus dem Jahr 2012 als Grundlage gewählt[30] und durch eigene Überlegungen und besonders relevant erscheinende Kriterien vervollständigt. Einige Kriterien, wie zum Beispiel die Marktakzeptanz, wurden dabei aus dem Bewertungsspektrum entfernt, da eine Bewertung der Marktakzeptanz nicht mit dem eigentlichen Ziel dieser Arbeit übereinstimmt. Nach Möglichkeit sollen diese Aspekte im allgemeinen Teil der jeweiligen Ausarbeitungen im Praxisteil dennoch vorgestellt werden, lediglich eine Bewertung soll nicht vorgenommen werden.

Sicherheitsqualität	**Funktionalität**	**Umsetzung**
Sicherheit	Funktionsumfang	Support
Zugriffsrechte	Konfigurierbarkeit	Benutzerfreundlichkeit
Ausfallsicherheit	Skalierbarkeit	Implementierungsaufwand
Logging	Kombinierbarkeit	Release Abstände
Fehlerrate	Performance	Dokumentation

Tabelle 1: Bewertungskriterieren sortiert unter Oberbegriff

Tabelle 1 zeigt die ausgewählten Kriterien sortiert in drei Hauptbereiche unter den Begriffen Sicherheitsqualität, Funktionalität und Umsetzung. Unter Sicherheitsqualität fallen die Punkte Sicherheit, Zugriffsrechte, Ausfallsicherheit, Logging und Fehlerrate. Funktionalität beinhaltet

[30] Vgl. Golembowska, A. u.a. (2012)

den Funktionsumfang, die Konfigurierbarkeit, die Skalierbarkeit, die Kombinierbarkeit und die Performance. Der dritte Hauptbereich ist die Umsetzung mit den Unterpunkten (laufender) Support, Benutzerfreundlichkeit, Implementierungsaufwand, Release-Abstände und Dokumentation. Die Überlegung hinter dieser Einteilung ist, dass in der Auswertung nicht eine einzige Tabelle mit 15 Kriterien entsteht, sondern drei verschiedene übersichtliche Tabellen, die man sowohl untereinander, als auch insgesamt vergleichen kann. Dies hat den großen Vorteil, den Fokus der Bewertung jederzeit auch als Leser selbst legen zu können sowie eine insgesamt tiefere Bewertung der Thematik zu erreichen. Des Weiteren ist es möglich, einen Fokus auf eines der drei Themen zu legen und diesen besonders hervorzuheben. Die einzelnen Unterpunkte werden in der Erläuterung des Bewertungsverfahrens noch genauer vorgestellt.

Um eine möglichst genaue und nachvollziehbare Bewertung zu erreichen, soll eine Nutzwertanalyse angewendet werden. Diese ermöglicht es, unter verschiedenen ausgewählten Kriterien unterschiedliche Lösungen zu analysieren und auf ihr Erfolgspotenzial zu bewerten. Dabei können sowohl quantitative wie auch qualitative Kriterien mit in die Bewertung einbezogen werden, um am Ende eine Rangfolge der verschiedenen Lösungen zu generieren. In der Nutzwertanalyse wird jedem Kriterium eine Gewichtung zugeteilt entsprechend der Relevanz für die Auswertung. Alle Kriterien werden danach bewertet und mit ihrer Gewichtung multipliziert. Die Ergebnisse werden dann zu einem Gesamtwert addiert für ein möglichst übersichtliches Ergebnis. K.O.-Kriterien erhalten so zum Beispiel die höchste Gewichtung und fließen stärker mit in die Gesamtbewertung ein. Dies ermöglicht es, eine Vielzahl von Möglichkeiten zu bewerten, ohne dabei den eigentlichen Fokus zu verlieren. [31]

Nutzwertanalyse	*Gewichtung*	*Produkt A*	*Produkt B*
Kriterium A	9	4	6
Kriterium B	3	8	4
Gesamt		60	66

Tabelle 2: Einfaches Beispiel Nutzwertanalyse

Tabelle 2 zeigt ein einfaches Beispiel einer Nutzwertanalyse mit einer Skala von 1 (sehr niedrig) bis 10 (sehr hoch). In den horizontalen Überschriften ist die Gewichtung gegeben, sowie das Produkt A und das Produkt B, eine Erweiterung um Produkt C bis X wäre jederzeit möglich, wenn benötigt. Auf der Seite der vertikalen Überschriften sind die Kriterien A und B gegeben, sowie das Gesamtergebnis. Auch hier könnte man den Kriterienkatalog nach Wunsch erweitern. Die Gewichtung im obigen Beispiel zeigt, dass Kriterium A mit 9 Punkten

[31] Vgl. Winter, S (2014), S.135ff

deutlich relevanter ist für Bestimmung des richtigen Produktes als Kriterium B. Dementsprechend ist es auch zu erklären, dass Produkt B insgesamt mit 66 Punkten die bevorzugte Lösung darstellt, obwohl es in der Bewertung der Kriterien mit sechs und vier Punkten schlechter aufgestellt scheint als Produkt A mit vier und acht Punkten.

Für das vorliegende Projekt wurde die Nutzwertanalyse den Anforderungen entsprechend angepasst. Es wird nicht eine große Tabelle mit 15 Kriterien geben, sondern wie bereits erwähnt drei verschiedene. Pro Kapitel wird nur ein Produkt bewertet und eine Zusammenführung der unterschiedlichen Ergebnisse findet am Ende in der Analyse der Ergebnisse statt. Als Skala für die Bewertung wurde eine Reichweite von eins bis drei Punkten gewählt. Aufgrund der teilweise schweren Bewertbarkeit einiger Kriterien können so sinnvoll Tendenzen bewertet werden von eins (niedrig) bis drei (hoch). Die Verfasser versprechen sich durch diese Bewertung von Tendenzen außerdem eine hohe Objektivität in der Bewertung. In Anlehnung daran wurde auch die Gewichtungsskala von eins bis zehn auf eins bis fünf angepasst, um die Ergebnisse nicht durch eine zu starke Gewichtung zu verfälschen. Zusätzlich wurde ein Faktor eingeführt, um die Bedeutung der jeweiligen Hauptgruppen in die Bewertung mit einfließen zu lassen. Die einzelnen Tabellen werden in den nachfolgenden Abschnitten nun kurz vorgestellt. Die Gedanken zu der gewählten Gewichtung beruhen dabei auf einem Gespräch mit dem Auftraggeber und dem Verständnis des Verfassers.[32]

Sicherheitsqualität	*Gewichtung*	*Bewertung*
Sicherheit	4,5	
Zugriffsrechte	5	
Ausfallsicherheit	4	
Logging	3	
Fehlerrate	3,5	
Gesamt	Faktor 2x	

Tabelle 3: Vorlage Nutzwertanalyse zur Sicherheitsqualität

Tabelle 3 veranschaulicht die Vorlage der Nutzwertanalyse für die Sicherheitsqualität. Insgesamt ist die Sicherheitsqualität das wichtigste Thema in dieser Auswertung und wurde folglich mit dem Faktor zwei belegt um ihre Wichtigkeit in der Gesamtauswertung weiter hervorzuheben. Das Kriterium Zugriffsrechte hat die höchste Gewichtung mit einem Wert von fünf. Wie aus den Gesprächen mit dem Auftraggeber (.Versicherung) hervorging, ist es ein sehr wichtiges Anliegen beim Betrachten von verschiedenen Möglichkeiten, dass Zugriffsrechte exakt und sicher vergeben werden können. Die Sicherheit ist mit 4,5 Punkten bewertet worden, da die innere Sicherheit der betrachteten Lösung sehr hoch seien muss, um eine Anwendung zu ermöglichen. Es folgt die Ausfallsicherheit mit 4 Punkten, diese ist von großer

[32] Vgl. Dittkowski, M. / Nitsche, U. / Schäfer, B. (2014)

Bedeutung, da ein Ausfall das tägliche Geschäft stören würde und erhebliche Konsequenzen hätte. Es folgen Fehlerrate mit 3,5 und Logging mit 3 Punkten. Diese sind in ihrer Gewichtung als mittelmäßig bis überdurchschnittlich anzusehen und bewerten, in wie weit die Richtigkeit und Vollständigkeit der Daten zu jedem Zeitpunkt gewährleistet werden kann.

Funktionalität	*Gewichtung*	*Bewertung*
Funktionsumfang	4,5	
Konfigurierbarkeit	3	
Skalierbarkeit	2,5	
Kombinierbarkeit	3,5	
Performance	4	
Gesamt	Faktor 1,5x	

Tabelle 4: Vorlage Nutzwertanalyse zur Funktionalität

Tabelle 4 zeigt die Vorlage der Nutzwertanalyse zum Thema der Funktionalität. Es wird ein Gesamtfaktor von 1,5 gewählt, da die Funktionalität zwar sehr wichtig ist, aber nicht an die Bedeutung der Sicherheitsqualität herankommt. Die wichtigsten Kriterien in der Funktionalität sind der Funktionsumfang mit 4,5 und die Performance mit 4. Zusammen garantieren diese beiden Punkten eine qualitativ und quantitativ hohe Funktionalität in den gewünschten Einsatzbereichen und haben damit die höchste Bedeutung in diesem Bereich. Die Kombinierbarkeit ist mit 3,5 bewertet worden, da eine Standalone Lösung aus Sicht der Verfasser unrealistisch erscheint und es dementsprechend wichtig ist, verschiedene Kombinationsmöglichkeiten zu haben. Die Skalierbarkeit mit 3 und Konfigurierbarkeit mit 2,5 hingegen sind in diesem Bereich von etwas untergeordneter Relevanz, da im zu untersuchendem Fall das System nach Möglichkeit nur einmal aufgesetzt werden und anschließend dauerhaft mit ausreichend Kapazität betrieben werden soll.

Umsetzung	*Gewichtung*	*Bewertung*
Support	3	
Benutzerfreundlichkeit	3,5	
Implementierungsaufwand	4	
Release Abstände	1,5	
Dokumentation	2,5	
Gesamt	Faktor 1x	

Tabelle 5: Vorlage Nutzwertanalyse zur Umsetzung

Tabelle 5 stellt die Vorlage der Nutzwertanalyse zur Umsetzung dar. Es wird der Faktor 1 gewählt, da im Gegensatz zur Sicherheitsqualität und Funktionalität die Möglichkeit der Umsetzung eine verhältnismäßig geringere Rolle spielt. Dies ist zu erklären, da ein Teil der Umsetzung einmalig ist und die tägliche Nutzung eher von den anderen Bereichen abhängt.

Dennoch ist der Implementierungsaufwand mit 4 hoch bewertet, da er innerhalb der Umsetzung einen sehr wichtigen Bereich einnimmt. Die Benutzerfreundlichkeit ist mit 3,5 bewertet. Diese ist zwar eigentlich enorm wichtig, da sich das Projekt aber auf einer Linux-Plattform bewegt, muss davon ausgegangen werden, dass die Benutzer die erforderlichen Grundkenntnisse besitzen und eventuell auch mit komplexeren Oberflächen umgehen können. Im Falle von Problemen ist es zusätzlich wichtig Hilfe vom zuständigen Supportteam zu erhalten, dies wird mit 3 gewichtet und ist ungefähr auf einem Niveau mit der Dokumentation (2,5), die ähnliche Ziele verfolgt. Den Release-Abständen wird mit 1,5 nur eine sehr geringe Wichtigkeit zugeteilt. Es ist nicht von Interesse wöchentlich Updates auf das System aufspielen zu müssen. Trotzdem ist eine Verbesserung der ursprünglichen Releases natürlich positiv.

Diese drei verkleinerten Nutzwertanalysen sollen in den nachfolgenden Kapiteln nun eingesetzt werden um die verschiedenen Möglichkeiten zu analysieren und zu bewerten. In der Schlussanalyse sollen die erarbeiteten Ergebnisse dann schrittweise zusammengeführt werden, um ein Urteil über eine etwaige Reihenfolge treffen zu können.

7 Access Control List (ACL)

Im Folgenden wird ein Überblick über die Funktionsweise der Access Control List gegeben und anschließend eine Bewertung unter den definierten Aspekten vorgenommen. Es ist zu unterscheiden zwischen den ACLs die auf Router- und Netzwerkkommunikation angewendet werden und die ACLs in Extended File Systems unter Linux. Bestandteil dieser Ausarbeitung sind nur die ACLs in Extended File Systems, da diese die Sicherheit der verwalteten Dateien im System gewährleisten und somit für das Ziel dieser Arbeit relevant sind.

7.1 ACL im Überblick

Unter der Access Control List versteht man eine Funktionalität unter Linux, die es dem Administrator erlaubt, Nutzerrechte an User oder Gruppen zu vergeben und so die Sicherheit der Daten zu gewährleisten. Im Allgemeinen wird zwischen 3 Rechten unterschieden: r (read), w (write) und x (execute). Wird einem User oder einer Gruppe kein Recht erteilt, so wird dies durch ein „-“ gekennzeichnet. Mit dem read-Recht darf der User oder die Gruppe das Objekt sehen aber nicht verändern. Im Gegensatz dazu steht das write-Recht, das dem User erlaubt, das Objekt zu verändern. Mit dem execute-Recht ist es dem User erlaubt, das Objekt auszuführen.[33]

Die ACL verfeinert diese Rechtevergabe noch und ermöglicht es, einzelnen Gruppen oder Usern verschiedene Rechte auf Dateien oder Verzeichnisse zu geben. So können beispielsweise User A und User B Mitglieder der Gruppe USERS sein, können jedoch unterschiedliche Berechtigungen für ein bestimmtes Objekt haben, USERS darf lesen, User A zusätzlich noch schreiben.[34]

Voraussetzung für den Gebrauch von ACLs unter Linux ist die Verwendung einer Kernelversion ab 2.5.46. Dieser unterstützt ACLs für die Dateisysteme EXT2, EXT3, XFS, JFS und ReiserFS. Grundsätzlich muss für jede Distribution geschaut werden, ob ACL direkt unterstützt werden.[35]

[33] Vgl. Suse (2003)
[34] Vgl. Novell (2011)
[35] Vgl. Suse (2003)

7.2 Funktionsweise von ACL

Die ACLs werden über Shell-Befehle (Befehle über die Kommandozeile ohne spezielle grafische Oberfläche) auf verschiedene Verzeichnisse oder Dateien angewandt. Grundlegend gibt es zwei Befehle: *getfacl* und *setfacl*.

Mit *getfacl* können die bereits angewandten ACLs auf eine Datei angezeigt werden und so die vergebenen Rechte ausgelesen werden. Als Ausgabe erscheint dann eine Liste von Einträgen mit Benutzern oder Gruppen und ihren spezifischen Rechten auf die angeforderte Datei oder das angeforderte Verzeichnis. Außerdem gibt es einen *mask*-Eintrag. Diese Maske legt fest, welche Rechte auf die Datei wirklich angewendet werden dürfen. Lässt die Maske nur Leserechte zu, ist es unerheblich, ob ein Benutzer auch Schreib- und Ausführungsrechte bekommen hat. Dies stellt einen zusätzlichen Sicherheitsmechanismus dar.[36]

Mit *setfacl* können die Rechte für Benutzer und Gruppen vergeben werden. Es wird immer zwischen Benutzern, Gruppen und Anderen unterschieden. Indem man beispielsweise Anderen keine Rechte für die zu schützende Datei gibt, ist diese vollständig abgeschirmt gegen den Zugriff von außen.[37]

Aufbau der Kommandos zum Setzen der ACLs:

- **setfacl -m <ACL-SPEC>**
  setzen / modifizieren von ACL
- **setfacl -M <Quelldatei> <Datei oder Ordner>**
  setzen / modifizieren von ACL
- **setfacl -x <ACL-SPEC> <Datei oder Ordner>**
  ACL entfernen
- **setfacl -X <Quelldatei> <Datei oder Ordner>**
  ACL entfernen
- **setfacl -b <Datei oder Ordner>**
  Alle ACL-Einträge entfernen
- **setfacl -k <Ordner>**
  Alle Default ACL-Einträge entfernen[38]

Setzt man sogenannte Default-ACLs, so werden diese beim Anlegen eines Unterverzeichnisses an dieses vererbt und regelt automatisch den Zugriff im neu angelegten Teil.[39]

[36] Vgl. ebenda
[37] Vgl. ebenda
[38] Vgl. Sarton, J. J. (o.J.):

7.3 Sicherheitsqualität

Sicherheit:

Die allgemeine Sicherheit unter Verwendung von ACLs ist in jedem Fall gewährleistet. Durch die verfeinerte Rechtevergabe und die Vererbung der ACLs an Unterverzeichnisse, sind die Dateien vor fremdem Zugriff sehr gut geschützt. Die allgemeine Sicherheit ist demnach hoch und wird mit 3 bewertet.

Zugriffsrechte:

Wie schon beschrieben, werden die Zugriffsrechte mit den ACLs optimal ergänzt und können so sehr filigran gesetzt werden. Es ist genau geregelt, welcher Benutzer oder welche Gruppe welches Recht auf die bestimmte Datei oder das Verzeichnis anwenden darf. Daher wird dieser Aspekt mit 3 bewertet.

Ausfallsicherheit:

Die Ausfallsicherheit der ACLs hängt von der Ausfallsicherheit des Betriebssystems bzw. des Filesystems ab, da die ACLs Bestandteil dessen sind. Zur Ausfallsicherheit lässt sich hier nur schwer eine Aussage treffen, da diese von der verwendeten Konfiguration bzw. der Hardware abhängig ist. Da grundsätzlich die Datenkonsistenz durch das Extended File System gewährleistet ist, wird dieser Punkt mit 2 bewertet.

Logging:

In den Extended File Systems wird, wie bereits beschrieben, jeder Schritt vorher in einem Journal gespeichert und kann somit nachvollzogen werden. Aus diesem Grund ist immer ein Überblick vorhanden welcher User in welchem Verzeichnis, welche Datei bearbeitet oder gelöscht hat. Dieses Journal übernimmt dieselbe Funktion wie ein Log.[40] Daher wird dieser Punkt mit 3 bewertet.

Fehlerrate

[39] Vgl. Suse (2003)
[40] Vgl. LinuxWiki (2013)

Zur Fehlerrate der ACLs lässt sich keine Aussage treffen, da nicht bekannt ist, ob und in welchem Maße Fehler anfallen. Vorausgesetzt, die Rechte auf eine Datei wurden korrekt konfiguriert, werden keine Fehler anfallen und nur die User Zugriff haben, denen die Rechte auch galten. Kommt es zur falschen Vergabe der Rechte auf eine Datei, so kann es zu falschen Zugriffen kommen. Dies liegt dann nicht an Fehlern im Code, sondern an individuellem Fehlverhalten.

Aus den Untersuchungen ergeben sich folgende Ergebnisse:

Sicherheitsqualität ACL	*Gewichtung*	*Bewertung*
Sicherheit	4,5	3
Zugriffsrechte	5	3
Ausfallsicherheit	4	2
Logging	3	3
Fehlerrate	3,5	NA
Gesamt	Faktor 2x	91

Tabelle 6: Nutzwertanalyse Sicherheitsqualität ACL

7.4 Funktionalität

In diesem Abschnitt werden die Funktionen der ACLs untersucht und bewertet.

Funktionsumfang:

Mit ACLs können die Benutzerrechte besonders präzise vergeben und auch wieder entzogen werden. Jedoch ist dies die einzige Funktion von ACLs. Aus diesem Grund wird dieser Punkt mit 1 bewertet. Allerdings muss dazu erwähnt werden, dass diese Funktion den Zweck der ACLs bereits voll erfüllt und zur Zugriffskontrolle nicht mehr benötigt wird.

Konfigurierbarkeit:

Die Konfigurierbarkeit der ACLs ist sehr hoch, da sie eine Verfeinerung der Standard-Rechtevergabe darstellt. Die drei Rechte können je nach Bedarf auf einzelne Dateien oder ganze Verzeichnisse oder sogar Dienste vergeben werden. Dabei kann man sie ganzen Gruppen oder nur einzelnen Benutzern zuweisen und zusätzlich mit Masken die effektive Rechtenutzung auf einem Verzeichnis oder einer Datei festlegen. Die Konfigurierbarkeit wird demnach mit 3 bewertet.

Skalierbarkeit:

Die Skalierbarkeit der ACLs ist durch die Extended File Systems definiert. Je nach Wahl des Filesystems können diese gut oder weniger gut auf große Speichermengen skalieren. Beispielsweise ist XFS ein Filesystem, das eine hohe Skalierbarkeit durch Einteilung des Speichers in gleich große Zuweisungsgruppen sicherstellt. Diese Zuweisungsgruppen können als eigene Dateisysteme im Dateisystem gesehen werden und gleichzeitig vom Kernel adressiert werden.[41] Aus dem Grund, dass die Skalierbarkeit eben abhängig von der Wahl des Dateisystems ist, wird dieser Aspekt mit 2 bewertet.

Kombinierbarkeit:

Die ACLs sind mit anderen Sicherheitsmechanismen zur Abschirmung des Kernels oder zur sicheren Netzwerkkommunikation kombinierbar. ACLs regeln nur den internen Filesystemzugriff, was zusätzliche Arten der Sicherheit auch notwendig macht. Da die ACLs standardmäßig Bestandteil des Linux-Kernels bzw. des Filesystems sind, können sie mit beliebigen von Linux unterstützter Software zur weiteren Absicherung des Gesamtsystems kombiniert werden.[42] Da die Kombinierbarkeit keinen Einschränkungen unterliegt wird sie mit 3 bewertet.

Performance:

Die Entwickler der Suse-Distribution haben sechs Extended Filesystems mit und ohne aktivierte ACLs getestet. Dabei wurde jeweils die Zugriffszeit auf eine Datei nach einem Reboot gemessen. Die Messung lieferte folgende Ergebnisse:

	Without ACL	With ACL
Ext2	9	1743
Ext3	10	3804
ReiserFS	9	6165
XFS-256	14	7531
XFS-512	14	14
JFS	13	13

Tabelle 7: Ergebnisse Performancetest[43]

Bei nur zwei der sechs Filesysteme ist die Zugriffszeit gleich geblieben, da diese die ACL-Informationen direkt im Filesystem speichern und so keine Leseoperation in einem anderen

[41] Vgl. Suse (2011)
[42] Vgl. Suse (2003)
[43] Vgl. Suse (2011)

Speicher ausführen müssen. Bei allen Anderen ist die Zugriffszeit erheblich höher. Durch diese Schwankung und die nur geringe Anzahl der Filesysteme, bei denen die Zugriffszeit gleich geblieben ist, wird dieser Aspekt mit 1 bewertet.

Funktionalität ACL	*Gewichtung*	*Bewertung*
Funktionsumfang	4,5	1
Konfigurierbarkeit	3	3
Skalierbarkeit	2,5	2
Kombinierbarkeit	3,5	3
Performance	4	1
Gesamt	Faktor 1,5x	49,5

Tabelle 8: Nutzwertanalyse Funktionalität ACL

7.5 Umsetzung

Im letzten Abschnitt werden allgemeine Aspekte zu Extended File Systems und ACLs untersucht.

Support:

Der Support der Filesysteme hängt von ihrem Alter und vom Stellenwert des jeweiligen Filesystems in der Linux-Welt ab. EXT3 und EXT4 werden durch die Entwickler weiterhin unterstützt. Filesystems wie JFS wurden von IBM entwickelt und werden daher auch von diesen unterstützt. Es gibt eigens angelegte Projekt-Webseiten, um dort den Entwicklern Fehler melden zu können und diese anschließend ausgebessern zu lassen.[44] Da einige aktuelle Releases schon ein paar Jahre alt sind, ist es schwierig den aktuellen Support einzuschätzen. Die Bewertung ergibt daher eine 2.

Benutzerfreundlichkeit:

Die ACLs werden über Shell-Kommandos gesetzt. Dies ist für erfahrene Benutzer leicht machbar. Außerdem ist auch eine grafische Oberfläche (Eiciel) zur Verwaltung der ACLs verfügbar, die separat installiert werden muss.[45] Damit können auch weniger erfahrene Benutzer die ACLs für die Dateien setzen. Aus dem Grund, dass diese aber erst separat installiert werden muss und standardmäßig nur Shell-Befehle gelten, wird dieser Aspekt mit 2 bewertet.

Implementierungsaufwand:

[44] Vgl. Linux Kernel Organization (2014a)
[45] Vgl. Ibáñez, R. (2014)

Der Implementierungsaufwand richtet sich nach dem Anwendungsfall. Sollen die Rechte für ein kleines Unternehmen verwaltet werden, so ist der Aufwand eher gering. Für ein sehr großes Unternehmen kann der Aufwand hingegen sehr groß ausfallen, da hier viele Verzeichnisse und Dateien vorhanden sind, auf die viele verschiedene Benutzer verschiedene Rechte haben. Durch die Einteilung der User in Gruppen, müssen nicht jedem Benutzer einzeln die Rechte zugewiesen werden. Außerdem ist die Vererbung von ACLs durch die Default-ACLs sehr hilfreich um Unterverzeichnisse automatisch mit denselben Rechten auszustatten, wie die des Oberverzeichnisses. Aufgrund dieser Erleichterungen wird der Aufwand verringert und so mit einer 2 bewertet.

Release Abstände:

Die Release-Abstände für die Werkzeuge der Dateisysteme EXT2, EXT3 und EXT4 sind von 5 Versionen pro Jahr (2009 und 2010) auf 3 Versionen pro Jahr gefallen. Das kann daran liegen, dass zu Beginn mehr Fehler im Code ausgebessert werden mussten und so mehrere Versionen nötig waren. Später sind es dann eher neue Funktionen, die hinzugefügt worden sind und eine neue Version erforderlich machten.[46] Für JFS ist ebenfalls die Release-Anzahl von 3 Releases im Jahr 2005 auf 1 Release in den Folgejahren bis 2011 gefallen.[47] Da die Release-Abstände wie zu erwarten größer werden, jedoch Dateisysteme wie JFS bereits seit 2011 keine Neuerung erfahren haben, wird dieses Kriterium mit 2 bewertet.

Dokumentation:

Die Dokumentation ist für jedes Extended File System sowie für die ACLs durch zahlreiche Wiki-Seiten und fundierte Anleitungen vorhanden. Auf den projekteigenen Webseiten können alle Informationen zur Inbetriebnahme und Fehlerbehebung eingesehen werden. Außerdem ist der Quellcode gut dokumentiert, da viele Entwickler an dieser Software arbeiten. Als Resultat wird die Dokumentation mit einer 3 bewertet.

Zusammenfassend ergibt sich für den Aspekt der Umsetzung folgendes Ergebnis:

[46] Vgl. Linux Kernel Organization (2014b)

[47] Vgl. Kleikamp, D. (2013)

Umsetzung ACL	*Gewichtung*	*Bewertung*
Support	3	2
Benutzerfreundlichkeit	3,5	2
Implementierungsaufwand	4	2
Release Abstände	1,5	2
Dokumentation	2,5	3
Gesamt	Faktor 1x	31,5

Tabelle 9: Nutzwertanalyse Umsetzung ACL

8 Samba

8.1 Samba im Überblick

Samba ist eine Open Source Software, die es u.a. Linux- und Unix-Betriebssystemen ermöglicht, ein Windows-Filesystem zu emulieren und somit dem User den Eindruck zu vermitteln, er würde sich auf einer Windows-Server-Umgebung befinden. Der Name Samba wurde dabei aufgrund des Standardprotokolls für Netzwerkfreigaben unter Windows Server Message Block/Common Internet Filesystem (SMB/CIFS), welches von Samba verwendet wird, gewählt.[48] Die Software steht unter der GPL-Lizenz[49]. Der Quellcode muss daher frei zugänglich sein, die Software muss geteilt werden dürfen, es muss erlaubt sein die Software nach eigenen Wünschen anzupassen und evtl. Änderungen müssen – sofern sie veröffentlicht werden – auch unter der GPL-Lizenz veröffentlicht werden.[50]

Samba wurde 1992 als Open Source Projekt veröffentlicht[51], wobei die aktuellste Version nach weit über 200 Updates und Patches[52] Version 4.1.15 ist und das ständig an der Software arbeitende Team um die 30 Mitarbeiter umfasst[53]. Damit ist es mithilfe von Samba möglich, andere Betriebssystemumgebungen in eine Windows Active Directory Umgebung zu integrieren. Dabei ist es konfigurationsabhängig und spielt für die Grundfunktionsweise von Samba keine Rolle, ob es als einzelner User in dieser Active Directory Umgebung oder als ganzer Domain Controller eingesetzt wird.[54] Dies wird auch im Firmenslogan „Samba, Opening Windows to a Wider World!"[55] deutlich. Das Ziel hinter dem Projekt Samba ist es demnach, die „Barrieren der Interoperabilität [bedingt durch die Nutzung von verschiedenen Betriebssystemen] zu beseitigen"[56]

8.2 Funktionsweise von Samba

Wie bereits erwähnt, ist es mit Samba möglich, eine Linux Umgebung durch Emulation von Windows-Filesystemen in eine Windows Umgebung zu integrieren und somit mithilfe des SMB/CIFS-Protokolls gemeinsame Freigaben von Daten zu betrachten. Dies ist in beide

[48] Vgl. Eggeling, T. (2014)
[49] Vgl. Samba (o.J.a)
[50] Vgl. Free Software Foundation (2014)
[51] Vgl. Samba (o.J.a)
[52] Vgl. Samba (o.J.b)
[53] Vgl. H., J. (o.J.)
[54] Vgl. Samba (o.J.a)
[55] H., J. (o.J.)
[56] Ebenda

Richtungen möglich. Vom Linux System aus können sowohl Daten für die Windows Umgebung freigegeben werden, als auch Dateien, die auf der Windows Umgebung selbst freigegeben wurden eingesehen und verwendet werden.[57]

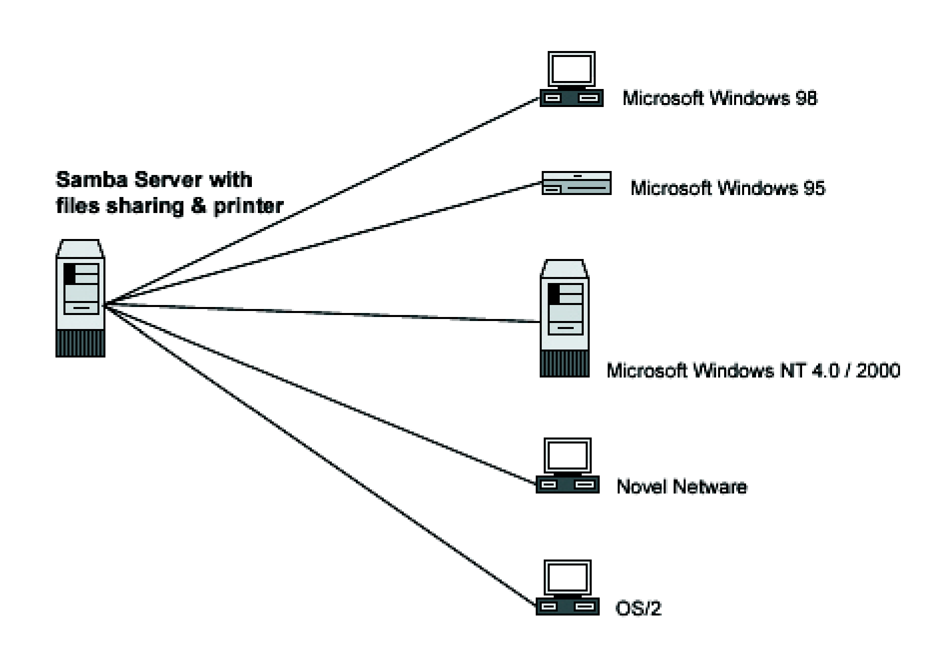


Abb. 2: Vernetzung von verschiedenen Betriebssystemumgebungen über einen File-Server mit Samba[58]

Wie in Abbildung 2 zu sehen, fungiert eine mit Samba konfigurierte Umgebung als Standardsystem in einem Netzwerk – in diesem Fall als Server für den Dateienaustausch und Druckdienste – unabhängig der anderen im Netzwerk im Einsatz befindlichen Betriebssysteme. Samba selbst stellt also über das SMB/CIFS-Protokoll eine für die anderen Systeme im Netzwerk verwendbare Schnittstelle zur Verfügung, um so eine barrierefreie Interoperabilität innerhalb des Netzwerks zu ermöglichen.

Der Zugriff der User auf das mit Samba betriebene Filesystem erfolgt nach erfolgreicher Installation und Konfiguration der Software, wie in anderen Systemen auch üblich, standardmäßig über den Benutzernamen und sein Passwort. Bei der Verwendung von Samba ist jedoch eine Schwierigkeit zu beachten. Samba verwendet eigene Benutzerkonten und somit auch eigene Passwörter.[59] Sofern bereits Linux-Benutzerkonten im System hinterlegt sind, ist bei einigen Linux-Distributionen eine Synchronisation dieser User-Daten möglich. Ist dies nicht der Fall, müssen die User für alle Personen, die Zugriff auf das mit Samba verbundene System benötigen, manuell erstellt werden, was besonders in größeren Unternehmen einen

[57] Vgl. Schmidt, M. (2006)
[58] Enthalten in: The Linux Documentation Project (o.J.)
[59] Vgl. Schmidt, M. (2006)

immensen Arbeitsaufwand benötigt. Eine Synchronisation mit den Benutzerdaten aus dem Windows Active Directory ist grundsätzlich nicht möglich.[60]

Die Rechtevergabe erfolgt bei Samba wie bei allen Linux-Distributionen mithilfe der UMASK-Regeln (Bsp.: „chmod 777“ als Befehl einen komplett öffentlich zugänglichen und veränderbaren Ordner/Datei). In einer Windows-Oberfläche lässt sich der mit Samba konfigurierte File-Server genau wie ein Windows-Server über die Option „Netzlaufwerk hinzufügen“ suchen und in den Explorer integrieren. Auch die Navigation innerhalb des File-Servers läuft dann so ab, wie in einem lokalen windowsbasierten Dateisystem.[61]

8.3 Sicherheitsqualität

Um eine Einschätzung über die Praktikabilität von Samba als Tool für die Unterstützung der Zugriffsicherheit wie eingangs beschrieben abgeben zu können, wird die Software nachfolgend mithilfe des bereits vorgestellten entwickelten Bewertungskatalogs analysiert. Die Analyse erfolgt wie bereits beschrieben hinsichtlich einzelner Charakteristika aus den Kategorien Sicherheitsqualität, Funktionalität und Umsetzung. Aufgrund der eigenen Spezifika die die Anforderungen dieses Seminararbeitsthemas mit sich bringen, erfolgt die Bewertung der Charakteristika basierend auf der Analyse des Verfassers als Gruppendiskussion.

Sicherheit:

Die ganzheitliche Sicherheit von Samba wird mit 2 Punkten bewertet. Diese Bewertung kommt dadurch zustande, dass die Sicherheit von Samba größtenteils von der Sicherheit des Firmennetzwerks im Allgemeinen abhängt. Die Software selbst bietet als Sicherheitsmechanismus die Notwendigkeit separater Benutzerkonten. Hier liegt es also am einsetzenden Unternehmen, Richtlinien – wie sie es auch für die normalen Arbeitssysteme geben sollte – festzulegen, um eine gewisse Passwortsicherheit der Benutzerkonten zu erreichen und zusätzlich das Netzwerk z.B. mit Firewalls zu schützen.

Zugriffsrechte:

Die Regelung der Zugriffsrechte unter Samba kann mit sehr gut bewertet werden. Dies liegt daran, dass der Zugriff auf die File-Server zum einen windowsseitig mithilfe der Nutzung des Active Directorys eingeschränkt werden kann und zum anderen an den eigenen Samba Be-

[60] Vgl. Eggeling, T. (2014)
[61] Vgl. ebenda

nutzerkonten, für welche die (für Linux üblichen) UMASK-Regeln unterschiedliche Rechte an verschiedene Benutzergruppen vergeben werden können.

Ausfallsicherheit:

Die Bewertung des Kriteriums Ausfallsicherheit erfolgt mit niedrig. Diese Herabstufung lässt sich mit der Funktionsweise von Samba begründen, da viele verschiedene Parteien an der Kommunikation beteiligt sind. Sollte also nur bei einer Partei von File-Server, Samba Software, SMB/CIFS-Protokoll und Endbenutzersystem ein Fehler auftreten, ist der Zugriff mindestens für den einzelnen Nutzer temporär nicht möglich.

Logging:

Grundsätzlich bietet der Zugriff auf Sambas Log-Dateien einige Konfigurationsmöglichkeiten beginnend mit dem Umfang. Problematisch ist jedoch die Tatsache, dass diese Log-Dateien äußerst komplex sind und nur von Personen mit Samba-Programmiererfahrung verstanden werden können, weshalb bei diesem Kriterium eine leichte Abstufung auf die Bewertung mittel vorgenommen wird.[62]

Fehlerrate:

Die Fehlerrate von Samba ist nicht bewertbar, da es keine Aussagen zum Fehlverhalten der Software aufgrund von Source Code-Fehlern gibt. Grundsätzlich lässt sich aber sagen, dass Fehler, welche die Performance von Samba verschlechtern wohl durch die große Community und verschiedenen Supportwege – hierauf wird im weiteren Verlauf noch eingegangen – relativ schnell behoben werden können.

Daraus ergibt sich folgendes Bild:

Sicherheitsqualität Samba	*Gewichtung*	*Bewertung*
Sicherheit	4,5	2
Zugriffsrechte	5	3
Ausfallsicherheit	4	1
Logging	3	2
(Fehlerrate)	3,5	NA
Gesamt	Faktor 2x	68

Tabelle 10: Nutzwertanalyse Sicherheitsqualität Samba

Die Sicherheitsqualität von Samba lässt sich, wie in Tabelle 10 zu sehen, als durchschnittlich beschreiben. Dies liegt vor Allem daran, dass Samba aufgrund seiner Funktionsweise nur

[62] Vgl. Eckstein, R. / Collier-Brown, D. / Kelly, P. (1999)

wenig bis gar keine eigenständigen Sicherheitsvorkehrungen bietet, sondern sich bereits in der Umgebung existierende Sicherheitsmaßnahmen zu Nutzen macht.

8.4 Funktionalität

Funktionsumfang:

Der Funktionsumfang wird mit 1 als eher niedrig bewertet, da das Programm grundsätzlich nur als Emulator einer Schnittstelle mit deren Zugriffsverwaltung dient. Weitere Zusatzfunktionen sind nicht vorhanden, wodurch der Umfang sehr gering ausfällt.

Konfigurierbarkeit:

Die Konfigurierbarkeit der Software wird als mittel eingestuft, da sie als Open Source Software unter der GPL-Lizenz mit frei verfügbarem Source Code erhältlich ist und an eigene Bedarfswünsche angepasst werden kann. Innerhalb der Software gibt es jedoch abgesehen von der Rechtevergabe keine weiteren Konfigurationsmöglichkeiten, da die Basiskonfiguration von den übrigen Netzwerkgegebenheiten abhängig und somit nur bedingt frei wählbar ist.

Skalierbarkeit:

Samba kann sowohl für kleine Heimnetzgruppen mit einem 1:1 Client-Server-Modell, als auch in großen Unternehmensnetzwerken eingesetzt werden. Somit bietet es dem Administrator eine größtmögliche Skalierbarkeit.

Kombinierbarkeit:

Auch die Kombinierbarkeit der Software ist sehr gut. Dies liegt zum einen daran, dass es aufgrund der Schnittstellenfunktion zwischen verschiedenen Systemen keine Einschränkungen auf Sicherheitssysteme wie Firewalls oder Sniffing-Tools lokal oder im Netzwerk ausübt und zum anderen an der Tatsache, dass mithilfe von Samba die Emulation von Windows-Filesystemen auf vielen anderen Betriebssystemen möglich ist.

Performance:

Die Performance der Software wird als mittel eingestuft. Sie hängt dabei hauptsächlich von der verwendeten Hardware ab. Außerdem ist das verwendete SMB/CIFS-Protokoll von der Performanceseite aus grundsätzlich mit anderen Filesharing-Protokollen vergleichbar.[63]

Daraus ergeben sich folgende Werte für die Nutzwertanalyse:

[63] Vgl. Dundee, P. / R., J. / H., J. (o.J.)

Funktionalität Samba	*Gewichtung*	*Bewertung*
Funktionsumfang	4,5	1
Konfigurierbarkeit	3	2
Skalierbarkeit	2,5	3
Kombinierbarkeit	3,5	3
Performance	4	2
Gesamt	Faktor 1,5x	54,75

Tabelle 11: Nutzwertanalyse Funktionalität Samba

Wie in Tabelle 11 zu sehen, kann das Kriterium Funktionalität von Samba durchschnittlich mit mittel bewertet werden. Das Spektrum geht dabei von einer eher schlechten Bewertung bis hin zu sehr guten Bewertungen von einzelnen Charakteristika, was sich durch die grundsätzliche Funktionsweise der Software erklären lässt.

8.5 Umsetzung

Support:

Der Support von Samba ist mit hoch zu bewerten, da es neben dem Mail-Support und Bug-Reporting auch eine globale Nachrichtengruppe und Mailinglist gibt. In dieser können u.a. Probleme von anderen Samba-Usern auf Basis ihrer Erfahrungen gelöst werden. Außerdem gibt es kommerziellen, länderspezifischen Sambasupport für derzeit 36 Länder.[64]

Benutzerfreundlichkeit:

Auch die Benutzerfreundlichkeit ist sehr hoch. Dies liegt zum einen daran, dass für die Konfiguration nur grundlegende Administrationskenntnisse von Netzwerken und Linux notwendig sind und zum anderen daran, dass den End Usern ein Windows-Filesystem vorgespielt wird. Somit wird es den Benutzern ermöglicht, in ihrer gewohnten Umgebung zu arbeiten.

Implementierungsaufwand:

Der Implementierungsaufwand von Samba ist vor allem im Vergleich zum Funktionsumfang sehr hoch, folglich mit niedrig zu bewerten. Dies ist zu begründen, da die Benutzerkonten und Rechte der einzelnen Benutzergruppen unter Umständen manuell eingegeben werden müssen, was zu einem enormen Mehraufwand führt. Bei bestimmten Voraussetzungen lassen diese sich jedoch – wie bereits erwähnt – mit Linux-Usern synchronisieren. Sind diese

[64] Vgl. Samba (o.J.c)

Voraussetzungen erfolgt, kann auch diese Eigenschaft mit sehr gut bewertet werden, da es sich bei der Implementierung dann lediglich um Installation und Einbindung in das Netzwerk handelt.

Release Abstände:

Neue Releases gibt es durchschnittlich alle 3 Wochen. Manchmal wird diese Zeit deutlich überschritten, häufig jedoch deutlich unterschritten was die hohe Bewertung dieser Eigenschaft ausmacht.[65]

Dokumentation:

Außerdem ist Samba sehr gut dokumentiert. Zu größeren Releases wie z.B. ab Version 3.5.x gibt es eigenständige Bücher die alle Schritte von Installation und Konfiguration in ihren Einzelheiten und Möglichkeiten detailliert beschreiben und somit als Step-by-Step-Handbuch und aufgrund der Beispiele auch zum Teil zum Bug-Fixing bzgl. Konfigurationsfehlern dienen können. Außerdem existiert ein eigenständiges Samba-Wiki in welchem die grundlegen Begriffe und Funktionen nochmals erläutert sind.

Umsetzung Samba	*Gewichtung*	*Bewertung*
Support	3	3
Benutzerfreundlichkeit	3,5	3
Implementierungsaufwand	4	1
Release Abstände	1,5	3
Dokumentation	2,5	3
Gesamt	Faktor 1x	35,5

Tabelle 12: Nutzwertanalyse Umsetzung Samba

Da, wie in Tabelle 12 ersichtlich, nur ein Kriterium abfällt, ist die Umsetzung von Samba mit sehr gut zu bewerten. Einzig das Kriterium Implementierungsaufwand fällt hier ab.

Zusammenfassend lässt sich feststellen, dass Samba eine gute bis sehr gute Möglichkeit darstellt, ein Windows-Filesystem in einer Linux Umgebung abzubilden und somit dem Endbenutzer eine Windows-Oberfläche vorzuspielen. Auch die Zugriffskontrolle lässt sich mit Samba aufgrund der eigenen Benutzerkonten realisieren. Des Weiteren ist es denkbar, logische Partitionen in der Linux Umgebung zu erstellen, welche dann nur bestimmten Benutzergruppen als Netzlaufwerke unter Windows zur Verfügung gestellt werden. Der einzige Nachteil an Samba besteht in den fehlenden eigenen Sicherheitsmechanismen. Abgesehen von einem separaten Benutzerkonto – welches u.U. sogar mit dem Linux User synchronisiert

[65] Vgl. Samba (o.J.b)

wird – bietet Samba keine eigenen Sicherheitsmechanismen und verlässt sich somit vollkommen auf bereits vorhandene Maßnahmen zur Netzwerksicherheit.

9 Linux Container (LXC)

9.1 LXC im Überblick

Linux Container, kurz LXC, ist ein Verfahren, was zur Virtualisierung auf Betriebssystemebene verwendet wird. Es ermöglicht die parallele Betreibung von mehreren voneinander isolierten Linux Systemen auf einem einzigen Linux-Host.[66] Hierbei handelt es sich nicht um virtuelle Maschinen, sondern um reine Prozessvirtualisierung.

Bei den Linux Containern handelt es sich um freie Software. Der Hauptteil des Codes ist unter der GNU LGPLv2.1+ Lizenz öffentlich zugänglich. Einige Android Komponenten sind unter der Standard „2-Klausel BSD“-Lizenz erhältlich, sowie einige Binärdateien und Templates unter der GNU GPLv2 Lizenz.[67]

Die GNU LGPLv2.1 (Lesser General License) Lizenz wurde von der Free Software Foundation entwickelt und mit einer Erstauflage (Version 2) 1991 veröffentlicht. Eine Modifizierung dieser Lizenz erfolgte 1999 unter dem Versionsnamen 2.1. Die aktuellste Version ist die Version 3.0, die im Jahre 2007 veröffentlicht worden ist. Die Bedingungen der GNU LGPLv2.1+ kann auf der Internetseite https://www.gnu.org/licenses/lgpl-2.1.html nachgelesen werden.[68]

Die „2-Klausel BSD“ Lizenz wurde 1998 von der Berkeley Universität in Kalifornien veröffentlicht.[69] Sie beinhaltet lediglich die zwei folgenden Klauseln und ermöglicht somit eine äußerst freie Nutzung des Source-Codes:

1. „Weiterverbreitete nichtkompilierte Exemplare müssen das obige Copyright, diese Liste der Bedingungen und den folgenden Haftungsausschluss im Quelltext enthalten.“[70]

2. „Weiterverbreitete kompilierte Exemplare müssen das obige Copyright, diese Liste der Bedingungen und den folgenden Haftungsausschluss in der Dokumentation und/oder anderen Materialien, die mit dem Exemplar verbreitet werden, enthalten.“[71]

Die GNU General Public License v2 (GPL) ist der Vorläufer der aktuellen GNU GPLv3. Der Inhalt ist über die folgenden Internetseite abrufbar: http://www.gnu.org/licenses/gpl-2.0.html

Die Idee der Container-Infrastruktur ist weit verbreitet und wird als Zukunftstechnologie angepriesen. Viele Unternehmen benutzen eine solche Infrastruktur bereits heute. Docker listet

[66] Vgl. Simon, C. (2014)
[67] Vgl. Linux Containers (o.J.)
[68] Vgl. Wikipedia (2014)
[69] Vgl. Open Source Initiative (o.J.b)
[70] Open Source Initiative (o.J.b)
[71] ebenda

Unternehmen wie Yelp, Bleacher Report, Spotify und ebay zu seinen Kunden.[72] Bei Docker handelt es sich jedoch nicht um eine reine LXC Struktur, vielmehr greift Docker Features wie cgroups und namespaces von dem Linux Kernel auf und erweitert diese um eigene Features.

Google setzt seit dem Jahr 2000 fast ausschließlich auf die Container-Architektur. Es ist die Rede von 2 Milliarden zusätzlichen Containern pro Woche, das entspricht 3300 Container in der Sekunde. Google startete diese Initiative nachdem die cgroups im Linux Kernel integriert worden sind.[73] Diese Initiative läuft unter dem Namen lmctfy, welches ausgeschrieben für „let me contain that for you" steht.[74]

Derzeit ist kein Unternehmen bekannt, welches die Linux Container in Verwendung hat. Dies kann jedoch durch die erst im Februar 2014 erfolgte Markteinführung erklärt werden.

Die vielseitigen Vorteile der Container sind die…[75]

- Senkung der Betriebskosten (keine Lizenzen nötig)
- Beschleunigung der Applikationsentwicklung
- Vereinfachung der Sicherheitseinstellungen
- Flexible Migration und Updates durch Container Clone
- Vereinfachung der Patcheinspielung (nur ein Betriebssystem muss gewartet werden)
- verbesserte Verteilung der Workloads (bessere Ressourcenausnutzung)
- Schnellere Bereitstellungszeit (VM=15 Minuten; Container=10 Sekunden)

9.2 Funktionsweise von LXC

Für die oben angesprochene Prozessvirtualisierung werden die Prozesse zu Gruppen zusammengeschlossen und in unterschiedliche Container geladen. Jedem Container werden eigene CPUs, Arbeitsspeichermodule und I/O Blöcke zugeordnet. Das wird durch das Cgroups Feature im Linux Kernel ermöglicht.[76] Die „[…] cgroups (Control Groups) dienen der Gruppierung von Prozessen. Dies ermöglicht dem Betriebssystem, einer Gruppe von defi-

[72] Vgl. Docker (2014)
[73] Vgl. Clark, J. (2014)
[74] Vgl. Marmol, V. / Jnagal, R. (2013)
[75] Vgl. Cisco (2014)
[76] Vgl. Archlinux (2015)

nierten Prozessen ausgewählte Ressourcen zuzuweisen.“[77] Hierbei werden zudem Hierarchien einzelner Prozesse auf den verschiedenen Ebenen berücksichtigt. Dabei gelten folgende Regeln:[78]

- jeder Prozess kann zur selben Zeit nur in einer Gruppe existieren
- neu erstellte Untergruppen sind leer
- eine Gruppe kann nur aufgelöst werden, wenn ihr keine Prozesse zugeordnet sind
- eine Gruppe kann nur aufgelöst werden, wenn keine weitere Untergruppe besteht

Ein weiteres Element der Linux Container sind die Namensräume (Namespaces), welche durch unterschiedliche Benennungen ein isoliertes System simulieren. Somit können alle Prozesse mit demselben Namensraum einander sehen, Prozesse mit einem anderen Namensraum jedoch nicht.[79] Eine Auflistung der durch Linux unterstützten Namensräume ist der Abbildung 3 zu entnehmen. Insgesamt gibt es 6 Namensräume für unterschiedliche Isolationstypen.

Namespace	Constant	Isolates
IPC	CLONE_NEWIPC	System V IPC, POSIX message queues
Network	CLONE_NEWNET	Network devices, stacks, ports, etc.
Mount	CLONE_NEWNS	Mount points
PID	CLONE_NEWPID	Process IDs
User	CLONE_NEWUSER	User and group IDs
UTS	CLONE_NEWUTS	Hostname and NIS domain name

Abb. 3: Unterstützte Namensräume in Linux

Außerdem unterstützt LXC chroot (change root) und Mandatory Access Control. Chroot kann dazu verwendet werden, dass ein Programm sein Wurzelverzeichnis wechselt. Nach einem Wechsel kann jedoch nicht mehr auf Dateien des alten Wurzelverzeichnisses zugegriffen werden.[80] Eine genauere Erklärung der Mandatory Access Control lässt sich unter dem Kapitel Zugriffsrechte finden.

Das LXC bietet ein Userspace-Interface mit dem die Sicherheitseinstellungen des Linux Kernels konfiguriert werden können. Eine umfangreiche und dadurch leistungsstarke API ermöglicht den Linux Usern im Zusammenspiel mit einfachen Tools eine einfache Handha-

[77] Gollub, D. / Seyfried, S. (2010)
[78] Vgl. ebenda
[79] Vgl. Kerrisk, M. (2015)
[80] Vgl. Linuxwiki (2011)

bung.[81] So können das System oder die Applikations-Container leicht erstellt und verwaltet werden.

Eine Abgrenzung zur ursprünglichen Virtualisierung kann der Abbildung 4 entnommen werden. Hier werden die oben beschriebenen Charakteristika noch einmal grafisch dargestellt. Sofort fällt auf, dass bei der Virtualisierung für jede Applikation (App A, App A',…) eine eigenständiges Guest OS erforderlich ist. Diese Ebene entfällt für die Linux Container vollständig. Diese zusätzliche Ebene sorgt dafür, dass es zu einem deutlichen Overhead (siehe Performance) und einer Ressourcenverschwendung kommt, aufgrund von Ressourcenbindung an die einzelnen Virtuellen Maschinen.

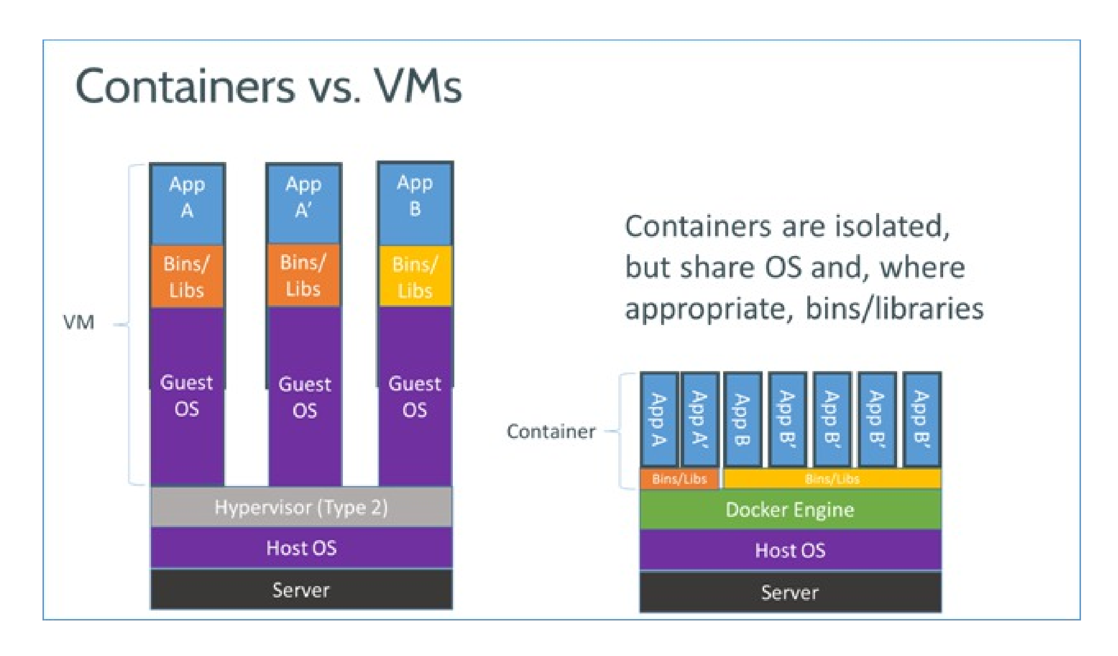


Abb. 4: Virtualisierung vs. Container[82]

9.3 Sicherheitsqualität

Im Folgenden wird die Sicherheitsqualität anhand der Kriterien allgemeine Sicherheit, Zugriffsrechte, Ausfallsicherheit, Logging und Fehlerrate bewertet.

<u>Sicherheit</u>:

Anfänglich konnte der Root-User des Containers Code auf dem Host-System ausführen.[83] Erst mit dem Release 1.0 wurden Namespaces eingeführt, die für mehr Sicherheit sorgen. Selbst wenn es dem Benutzer gelingt aus dem Container auszubrechen, so kann er trotzdem nicht als Root-User auf dem Host-System agieren, lediglich als normaler Benutzer. Zusätz-

[81] Vgl. Linux Containers (o.J.)
[82] Enthalten in: http://scm.zoomquiet.io/data/20131004215734/docker_vm.jpg
[83] Vgl. Debian (2011)

lich muss sichergestellt werden, dass die Konfiguration der einzelnen Container ein Übergreifen verhindert. [84]

Auch Cisco beschreibt diese Problematik. Red Hat verfolgt die Strategie, Container in Kombination mit SELinux als Basis zu betreiben. So soll der Host und weitere Container vor ungewollten Zugriffen geschützt werden. Das verwandte Projekt libseccomp erlaubt es dem Nutzer Systemaufrufe (syscalls) zu verhindern und so eine Attacke gegen den Host durch einen kompromittierten Container abzuwenden.[85]

Zugriffsrechte:

Für die Sicherung des Linux-System bezüglich der Schutzziele Vertraulichkeit, Integrität und Verfügbarkeit, bietet das Mandatory Access Control, kurz MAC, eine zusätzliche Schutzfunktion zu den Standard-Linux-Mechanismen.[86] Diese standardmäßige Rechteverwaltung implementiert zusätzlich die Discretionary Access Control (DAC). Hierbei werden die Rechte anhand der Identität eines Akteurs zugewiesen. Dazu wird zwischen Subjekten und Objekten differenziert. Beispielsweise sind Prozesse oder Anwender der Akteur und somit das Subjekt, Dateien dienen als Objekt. Im Folgeschritt vergleicht das DAC die Beziehung zwischen Subjekt und Objekt und entscheidet anhand der im System gespeicherten Relation, ob die Aktion durchgeführt wird oder nicht.[87] Während DAC Relationen kontrolliert, weist MAC via type-enforcement Typen zu, die alle Rechte verwalten.

Ein Sicherheitsproblem kann die Passwortverwaltung der einzelnen Benutzer darstellen, denn für das Ändern des Passwortes wird ein Schreibrecht unter Linux benötigt. Erhält der Benutzer ein Schreibrecht, so besteht die reelle Gefahr, dass dieser Passwörter von anderen Nutzern modifiziert.[88]

Ausfallsicherheit:

Ein Nachteil der LXC Infrastruktur ist die bedingte Ausfallsicherheit. Sollte ein fehlerhafter Container zu einem Absturz des Kernels führen, so werden auch alle anderen Container, die derzeitig auf diesem Kernel betrieben werden, abstürzen.[89] Der Kernel agiert hier nicht isoliert wie beispielsweise bei der klassischen Virtualisierung, bei der jede virtuelle Maschine einen eigenen Kernel besitzt.

[84] Vgl. Graber, S. (2014)
[85] Vgl. Cisco (2014)
[86] Vgl. Simon, C. (2014), S.47
[87] Vgl. ebenda
[88] Vgl. ebenda
[89] Vgl. Grimmer, L. (2013)

Jedoch leitet sich hierbei auch der große Vorteil der LXC Infrastruktur ab, da nur ein Kernel zur deutlichen Reduktion des Overheads führt (siehe Performance).

Logging:

Ein Fehlerlog ist über die Command-Zeile abrufbar.

Ein Aktionslog der Benutzer ist derzeit nicht in den Standardfeatures der Linux Container enthalten.

Fehlerrate:

Eine genaue Bestimmung der Fehlerrate ist für die Linux Container nicht möglich. Eine Fehlerrate kann nur durch den Absturz oder die Kompromittierung des Kernels erfolgen.

Daraus lässt sich die folgende Bewertung für die Sicherheitsqualität ableiten:

Sicherheitsqualität LXC	*Gewichtung*	*Bewertung*
Sicherheit	4,5	2
Zugriffsrechte	5	2
Ausfallsicherheit	4	2
Logging	3	1
(Fehlerrate)	3,5	NA
Gesamt	Faktor 2x	60

Tabelle 13: Nutzwertanalyse Sicherheitsqualität LXC

9.4 Funktionalität

In den nächsten Absätzen soll die Funktionalität der Linux Container anhand des Funktionsumfangs, der Konfigurierbarkeit, Skalierbarkeit, Kombinierbarkeit und der Performance gemessen werden.

Funktionsumfang:

Auf der offiziellen Seite der Linux Container sind folgende Kernel-Features gelistet: Die oben genannten Namespaces, Apparmor (Sicherheitsframework) und SELinux Profile, Seccomp policies (erlaubt einen Sandboxing Mechanismus), das ebenfalls oben beschriebene Filesystem chroots, weitere Kernel-Fähigkeiten sowie cgroups. Die Linux Container werden häufig als etwas zwischen chroot und vollwertiger Virtualisierung beschrieben. Das vorwiegende Ziel von LXC ist jedoch eine Umgebung zu erschaffen, die möglichst der Standard Linux Installation ähnelt, ohne weitere Kernel zu benötigen.[90]

[90] Vgl. Linux Containers (o.J.)

Konfigurierbarkeit:

Der Konfigurierbarkeit sind auf den ersten Blick keine Grenzen gesetzt. Kleinere Probleme wie die Passwortverwaltung der Benutzer können durch weitere Open Source Bausteine eliminiert werden. Um weitere Grenzen und Probleme aufzeigen zu können, müsste der Anwendungsfall im Detail betrachtet werden.

Skalierbarkeit:

RedHat gibt die Maximalanzahl von Containern auf einem Server mit 6000 an, zusätzlich sind 120 000 „Bind-Mounts" von Wurzelverzeichnissen im Dateisystem möglich. Bei diesen Angaben handelt es sich jedoch um die theoretische Skalierbarkeit. Daher gibt RedHat zu bedenken, dass praktisch gesehen alle Container parallel arbeiten, was zu einer Auslastung bzw. Überlastung des Systems führen kann, wenn alle Container gleichzeitig genutzt werden.[91]

Kombinierbarkeit:

Linux Container sind frei kombinierbar. Eine mögliche Kombination ist das Aufsetzen der Linux Container auf SELinux für eine gesicherte Handhabung durch den gehärteten Kernel. Des Weiteren werden durch Linux Container nur eine Vielzahl von eigenständigen Linux Systemen emuliert, auf denen auch weitere Sicherheitsmechanismen oder Software implementiert werden können.

Performance:

Es gibt verschiedene Möglichkeiten die Performance für LXC durch den Einsatz von cgroups zu verbessern. So kann beispielsweise eine dedizierte Verteilung der Maximalwerte von Hardwareressourcen für jeden einzelnen Container bestimmt werden. Zusätzlich dienen andere Tools zur Behebung von I/O Problematiken und führen somit zu einer hohen Transferrate.[92]

Generell sind die Performance von KVM (Virtual Machines) und Linux Containern annähernd gleich, da die KVMs in den vergangenen Jahren stark verbessert worden sind. Der Overhead ist jedoch deutlich geringer bei Linux Containern.[93]

[91] Vgl. Sarathy, B. / Walsh, D. (2013)
[92] Vgl. Xavier, Miguel G. u.a. (o.J.)
[93] Vgl. Lowe, S. (2013)

Daraus ergibt sich folgendes Bewertungsbild für LXC:

Funktionalität LXC	*Gewichtung*	*Bewertung*
Funktionsumfang	4,5	2
Konfigurierbarkeit	3	3
Skalierbarkeit	2,5	3
Kombinierbarkeit	3,5	2
Performance	4	3
Gesamt	Faktor 1,5x	66,75

Tabelle 14: Nutzwertanalyse Funktionalität LXC

9.5 Umsetzung

Für die Bewertung der Umsetzung von Linux Containern werden die Bereiche Support, Benutzerfreundlichkeit, Implementierungsaufwand, Release Abstände und die Dokumentation im Folgenden bewertet.

Support:

Der LXC Support ist abhängig von der weiteren Lebensdauer von Linux und deren Zusage, Verbesserungen und Sicherheitsupdates in regelmäßigen Abständen zu veröffentlichen. Es ist somit davon auszugehen, dass es einen Langzeitsupport für LXC geben wird.[94]

Zurzeit gibt es ein großes Release von LXC auf dem Markt, LXC 1.0, welches im Februar 2014 erschienen ist. Es wird bis April 2019 (5 Jahre) unterstützt werden. Grund dafür ist die Übernahme von LXC in das Ubuntu System 14.04, welches einen Langzeitsupport genießt.[95]

Benutzerfreundlichkeit:

Der Benutzer kann die Installation, Konfiguration und Verwaltung über die Command-Zeile realisieren. Eine weitere Möglichkeit ist die Nutzung von Hilfssoftware, die dem Benutzer eine grafische Oberfläche bietet, beispielsweise Docker.

Implementierungsaufwand:

Der Implementierungsaufwand einer LXC Umgebung ist in kleinem Rahmen überschaubar. Verschiedene Templates und Standardkonfigurationen unterstützen den Benutzer bei einer ersten Inbetriebnahme. Sollten die Container für einen gewissen Anwendungsfall identisch

[94] Vgl. Linux Containers (o.J.)
[95] Vgl. ebenda

sein, so reicht es, wenn der Benutzer einen einzigen Container konfiguriert und diesen im Anschluss klont.[96]

Für größere und komplexere Architekturen sowie für viele verschiedene Prozessgruppen eignet sich die Verwendung eines unterstützenden Programms wie beispielsweise Docker. Docker ermöglicht die variable Erstellung und dynamische Verwaltung einer solchen Umgebung.[97]

Ein erster Ansatz für die Installation kann aus dem Installationshandbuch von der offiziellen Linux Container Website unter dem Link (https://linuxcontainers.org/lxc/documentation/) oder der Ubuntu Website (https://help.ubuntu.com/12.04/serverguide/lxc.html) entnommen werden.

Release-Abstände:

Das erste Release LXC 1.0 erfolgte am 6. August 2008. Für Produktivumgebungen ist die Release Version LXC 1.0.0 6 Jahre später seit dem 20. Februar 2014 erhältlich. Das aktuellste Release ist LXC 1.0.7, wobei derzeit LXC 1.0.5 als stabiles Release gilt.[98]

Somit sind 7 Releases innerhalb der ersten 11 Monate nach Veröffentlichung der Linux Container für Produktivumgebungen erschienen. Das ergibt einen Durchschnitt von 0,63 Releases pro Monat. Die untenstehende Tabelle zeigt die tatsächlichen Daten der Veröffentlichung jedes einzelnen Releasestandes.

Release	Datum
LXC 1.0.0	20. Februar 2014
LXC 1.0.1	6. März 2014
LXC 1.0.2	27. März 2014
LXC 1.0.3	8. April 2014
LXC 1.0.4	13. Juni 2014
LXC 1.0.5	14. Juli 2014
LXC 1.0.6	24. September 2014
LXC 1.0.7	5. Dezember 2014

Tabelle 15: Veröffentlichung der Releasestände nach Datum

[96] Ubuntu (2014)
[97] Docker (2014)
[98] Vgl. Linux Containers (o.J.)

Dokumentation:

Eine vollständige Dokumentation, sowie das Installationshandbuch der Linux Container kann auf der Ubuntu-Website https://help.ubuntu.com/12.04/serverguide/lxc.html gefunden werden. Hier werden zudem die Charakteristika dieser Prozess-Virtualisierung kompakt dargelegt. Für den Einstieg ist diese Dokumentation empfehlenswert.

Daraus ergeben sich folgende Bewertungen der Nutzwertanalyse:

Umsetzung LXC	*Gewichtung*	*Bewertung*
Support	3	3
Benutzerfreundlichkeit	3,5	2
Implementierungsaufwand	4	2
Release Abstände	1,5	3
Dokumentation	2,5	3
Gesamt	Faktor 1x	36

Tabelle 16: Nutzwertanalyse Umsetzung LXC

10 Lightweight Directory Access Protocol (LDAP)

In diesem Kapital soll näher auf den Verzeichnisdienst LDAP eingegangen werden. Dabei wird zunächst eine allgemeine Übersicht über LDAP und seine Funktionalitäten gegeben, bevor auf das spezifischere OpenLDAP eingegangen wird. OpenLDAP ist dabei die für diese Arbeit relevante Lösung, die anschließend auf Sicherheitsqualität, Funktionalität und Umsetzung in den einzelnen Unterkapiteln analysiert werden soll. Im Verlauf der Bewertung in den Unterkapiteln ist OpenLDAP gemeint wenn von LDAP gesprochen wird.

10.1 LDAP im Überblick

LDAP, kurz für Lightweight Directory Access Protocol, ist ein Protokoll, das einen Verzeichnisdienst (Directory) unterstützt. Dies bedeutet, mit Hilfe von LDAP kann auf bestimmte Verzeichnisse zugegriffen werden. Dabei ist es für die Kommunikation zwischen dem Client und dem X.500-Verzeichnisdienst entwickelt worden, kann aber auch über einen LDAP-Server direkt auf das Verzeichnis zugreifen.[99] Verzeichnisdienste werden benötigt, da durch das stetige Wachstum die Ressourcen im Internet weit verteilt sind und teilweise redundant vorliegen. Ein einfaches Beispiel wäre der Zugriff auf verschiedene Netzwerke wie LAN, Universitätsnetzwerk und firmeninternes Intranet. Um einen zentralen Zugriff zu gewähren und gleichzeitig nur autorisierte Zugriffe zu gestatten, gibt es Verzeichnisdienste. Einige bekannte Verzeichnisdienste sind zum Beispiel DNS, NIS, Whois und X.500. Ein Verzeichnis ist eine Sammlung von Informationen über gewissen Objekten mit einer vorgeschriebenen Ordnung. Zu jedem Objekt ist es möglich, Detailinformationen abzurufen. Das System ist ähnlich einer relationalen Datenbank, unterscheidet sich aber in der Funktionalität. Ein Verzeichnisdienst spezialisiert sich auf sehr speziell konfigurierbare Suchanfragen und ist für den Lesezugriff optimiert. Schreibzugriffe sind oft stark limitiert. Dadurch ergibt sich die Hauptfunktionalität zum Auslesen von Daten, die nicht häufig geändert werden müssen. Einträge in Verzeichnissen wie LDAP werden als Entries gespeichert und können Attribute beinhalten. Die grundlegende Datenstruktur wird dabei als Schema bezeichnet und kann beliebig erweitert werden. Dabei müssen die Verzeichnisdienste nicht zwingend eine sichere Transaktion in Kauf nehmen. Da es sich nur um Lesezugriffe handelt, können kleinere Anomalien in Kauf genommen werden.[100]

LDAP ist ein einzelner Verzeichnisdienst, der dabei helfen soll, Daten über einzelne Personen einer Organisation abzurufen und zu speichern. Dabei kann ein großes Spektrum an

[99] Vgl. Schnabel, P. (2014)
[100] Vgl. Haberer, P (o.J.)

Informationen abgerufen werden. Durch das Universaldesign nach IETF können aus einem Schema viele spezialisierte Schemata aufgebaut werden. Zusätzlich ist es ein simples Protokoll, d.h. es ist einfach zu implementieren und wird von vielen Programmiersprachen, sowie Betriebssystemen unterstützt. Weiterhin erlaubt LDAP eine verteilte Architektur, durch die es möglich ist, Teile von LDAP-Servern auf physisch getrennten Rechnern zu speichern. Diese können über sogenannte Referrals miteinander kommunizieren. Durch diese Eigenschaften ermöglicht LDAP eine normalisierte Datenhaltung, eine Datenhaltung ohne Redundanzen. Außerdem ermöglicht es die zentrale Verwaltung der Informationen und gewährt Konsistenz in der Schnittstelle zum User, den Richtlinien für das Netzwerkmanagement und den Security Policies.[101]

Die Variante OpenLDAP ist, wie der Name bereits suggeriert, eine lizenzfreie Version von LDAP, die für die Allgemeinheit zugänglich ist und in der Netzgemeinschaft verbessert wird. Gestartet wurde OpenLDAP von der University of Michigan, beziehungsweise basiert es auf deren erstem Entwurf. Momentan ist die Version 2.4 auf dem Markt. Diese wird besonders in Linux Betriebssystemen verwendet, ist aber auch in Windows integriert, sowie für viele Programmiersprachen verfügbar. Die Version 3 ist bereits in Planung und soll in möglichst naher Zukunft verfügbar gemacht werden.[102] OpenLDAP ist dabei bereits viel mehr als nur noch ein simples Protokoll. Das Entwicklerteam spricht von der OpenLDAP Suite. Diese beinhaltet sowohl Entwicklerwerkzeuge, wie auch kleinere Applikationen. Beispiele wären ein Standalone LDAP Server oder verschiedene Libraries. Das Ziel des OpenLDAP Projektes ist es, eine robuste, wirtschaftliche und voll funktionale Open Source Lösung für LDAP zu Verfügung zu stellen.[103]

Von der Funktionalität gleicht OpenLDAP dabei 1 zu 1 der kommerziellen Version von LDAP. Beispiele der Anwendung von OpenLDAP ist eine zentrale Benutzerverwaltung oder ein Addressbuch. Es gibt den LDAP-Verzeichnisdienst der eine hierarchische Datenbank mit strukturierten Objekten beinhaltet. Diese können mit vielzähligen Attributen versehen werden. Objekte gehören zu einer Objektklasse, die eine Reihe von verschiedenen Attributen beinhaltet, um einen Eintrag in das Verzeichnis zu beschreiben. Dabei gibt es sowohl vordefinierte Klassen, wie auch die Möglichkeit eigene Vorstellungen umzusetzen. Ein Objekt ist dabei ein Distinguished Name, also ein eindeutiger Name, der durch seine Attribute einmalig definiert ist. Für die Installation von OpenLDAP muss zunächst die aktuelle Version heruntergeladen werden. Um diese vollständig kompilieren und mit LDAP kommunizieren zu kön-

[101] Vgl. Haberer, P (o.J.)
[102] Vgl. Schwaberow, V. (2001)
[103] Vgl. OpenLDAP Foundation (2014)

nen wird eine LDBM kompatible Datenbank benötigt, wie Berkeley DB2 oder GDBM. Diese sind in den meisten Distributionen enthalten, aber auch andere Back-Ends sind nutzbar.[104]

10.2 Funktionsweise von LDAP

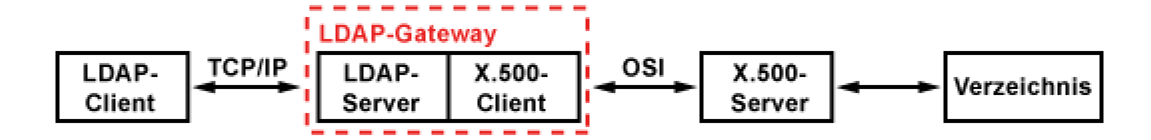


Abb. 5: Ablauf LDAP-Protokoll[105]

Abbildung 5 zeigt den typischen Ablauf des Protokolls. Der LDAP-Client greift auf den LDAP-Server über TCP/IP zu. Dieser LDAP-Server gehört zu einem LDAP-Gateway, zu dem auch ein X.500-Client gehört. Dieser übernimmt den Zugriff und greift über ein OSI-Protokoll-Stack auf einen X.500-Server zu, der eine direkte Verbindung zu dem gewünschten Verzeichnis besitzt. Es ist zu erkennen, dass es sich mit LDAP um ein simples Leseprotokoll handelt.

10.3 Sicherheitsqualität

Die Betrachtung des Themas Sicherheitsqualität bei LDAP gestaltet sich komplex. Dies liegt daran, dass einerseits LDAP selbst eigene Sicherheitsmechanismen vorweist, andererseits aber viele andere Sicherheitsmechanismen ebenfalls implementieren kann um die Sicherheitsqualität zu erhöhen. Für diese Ausarbeitung wird von einer OpenLDAP Version mit einer standardmäßigen Verankerung ausgegangen. Es wird nach Möglichkeit aber auf weitere Optionen eingegangen bzw. die zukünftige Bewertung in Klammern angegeben.

Sicherheit:

Der Punkt Sicherheit ist mit mittel (bis hoch) zu bewerten. Um auf den LDAP Service zugreifen zu können muss sich der Client erst Authentifizieren. Es gibt eine allgemeine Zugriffskontrolle. Diese ist in Version 2.x allerdings noch variabel und kann ein simples unverschlüsseltes Passwort sein, allerdings auch bereits eine verschlüsselte Kommunikation beinhalten. Dazu unterstützt LDAP das Simple Authentication and Security Layer (SASL) authentication framework. Außerdem wird Transport Layer Security (TLS) verwendet, um den sicheren Zu-

[104] Vgl. Schwaberow, V. (2001)
[105] Vgl. Schnabel, P. (2014)

gang zu gewährleisten.[106] Eine hohe Bewertung würde erfolgen, wenn man den Sicherheitsmechanismus ACL verwenden würde. Der genaue Problemfall ist im nächsten Abschnitt beschrieben.

Zugriffsrechte:

Der Punkt Zugriffsrechte ist als niedrig (bis hoch) zu bewerten. Dies liegt daran, dass in naher Zukunft zwar das ACL-Konzept standardmäßig implementiert werden soll und damit eine sehr gute Zugriffskontrolle gewährleistet wäre, momentan dies aber kein fester Teil der LDAP Spezifikation ist. Das ACL-Konzept würde es ermöglichen, den Zugriff nach einer Liste mit Zugangsrechten sehr detailliert festzulegen und die erlaubten Operationen zu definieren. Einige LDAP Implementierungen benutzen dies zwar bereits, allerdings in unterschiedlicher Weise und nicht in der Form um es hier als bewertbar einzuordnen.[107]

Ausfallsicherheit:

Obwohl es sich mit LDAP um ein einfaches Protokoll handelt, ist die Ausfallsicherheit in diesem Fall überraschend einfach zu bewerten. Diese ist sehr gut, da LDAP eine Replizierung der Daten auf verschiedene Rechner und Server ermöglicht. Fällt der Masterserver aus kann einfach mit dem Replica-Host weitergearbeitet werden. Des Weiteren findet so eine Verteilung der Last auf mehrere Server statt, da die vorhandenen Daten identisch sind und es sich um Lesezugriffe handelt.[108]

Logging:

Logging ist ein Punkt der unter LDAP nicht von erhöhter Bedeutung ist. Es erscheint sinnvoller, die Daten erneut vom Server abzufragen (on the fly), als auf einen gespeicherten Eintrag zurückzugreifen. Dennoch gibt es mit LDAP natürlich die Möglichkeit Log-Dateien anzulegen und sogar zu konfigurieren, was die Speicherung alles beinhalten soll. In wie weit man als Client Speicherplatz dafür belegen will, ist dabei dem Nutzer überlassen. Im Rahmen dieser Ausarbeitung werden die verschiedenen Logging Möglichkeiten unter LDAP nicht weiter untersucht und aufgrund der Vielfalt als hoch bewertet.[109]

Fehlerrate:

Die Fehlerrate bei LDAP liegt eher beim Nutzer als bei dem Protokoll selbst. Trotzdem ist Sie nicht zu vernachlässigen. LDAP ist als reines Lesetool bereit, kleinere Anomalien und Inkonsistenzen in Kauf zu nehmen, da diese dem Verzeichnis nicht schaden. Dies würde norma-

[106] Vgl. Haberer, P (o.J.)
[107] Vgl. ebenda
[108] Vgl. Banning, J. (2001)
[109] Vgl. ZyTrax Inc. (2015)

lerweise zu einer niedrigen Bewertung führen. Da aber hinter LDAP Implementierungen in den meisten Fällen transaktionssichere Datenbanken verwendet werden sind Fehler eigentlich nicht möglich. Aufgrund der Tatsache, dass LDAP alleine zu Fehlern neigt, aber mit transaktionssicheren Datenbanken keine Fehler möglich sind, ist dieses Kriterium mit mittel zu bewerten.

Insgesamt ergibt sich für die Sicherheitsqualität damit folgendes Bild:

Sicherheitsqualität LDAP	*Gewichtung*	*Bewertung*
Sicherheit	4,5	2 (3)
Zugriffsrechte	5	1 (3)
Ausfallsicherheit	4	3
Logging	3	3
(Fehlerrate)	3,5	2
Gesamt	Faktor 2x	84 (113)

Tabelle 17: Nutzwertanalyse Sicherheitsqualität LDAP

Tabelle 17 zeigt die Nutzwertanalyse für die Sicherheitsqualität von LDAP. Die vorher vorgestellten Bewertungen wurden eingetragen, mit der Gewichtung multipliziert und anschließend das Gesamtergebnis mit dem Faktor 2 verrechnet. Das Ergebnis sind 84 Punkte. Wie dies einzuschätzen ist, wird später im Gesamtvergleich herausgearbeitet. An dieser Stelle sollte festgehalten werden, dass die Sicherheitsqualität von LDAP vernünftig erscheint und mit den zukünftig geplanten Änderungen sehr vielversprechend wirkt. Der Verfasser möchte aber auch darauf hinweisen, dass vertieftes Wissen, zum Beispiel über die Sicherheit verschiedener Verschlüsselungsmechanismen, nicht vorhanden ist und deswegen das Vorhandenseins eines solchen bereits für die Höchstnote in obiger Bewertung ausreicht.

10.4 Funktionalität

Die Funktionalität von LDAP lässt sich deutlich einfacher bewerten als die Sicherheitsqualität. Das liegt besonders daran, dass, mit Ausnahme des Funktionsumfangs, die Kriterien direkt auf LDAP anwendbar sind und keinerlei Zusatzimplementierungen benötigt werden. Dennoch ist es an dieser Stelle wichtig darauf hinzuweisen, dass bei der Bewertung der anderen vier Kriterien nur LDAP auf seine Funktionalitäten überprüft werden kann.

Funktionsumfang:

Der Funktionsumfang von LDAP ist ursprünglich sehr schlank. Das bloße Auslesen von Daten aus einem Verzeichnis ist kein sonderlich großer Funktionsumfang. Doch durch die zusätzlichen Komponenten aus der OpenLDAP Suite, wie etwa einem LDAP-Server oder ver-

schiedenen Libraries wird der Umfang deutlich vergrößert.[110] Dennoch fehlen einige Bausteine um das gewünschte Gesamtergebnis mit dem momentanen Funktionsumfang von LDAP abzudecken. Eine mittlere Bewertung mit 2 Punkten ist angemessen.

Konfigurierbarkeit:

Die Konfigurierbarkeit wird mit 3 Punkten, also einer hohen Bewertung benotet. Diese Note setzt sich aus den endlosen Möglichkeiten zusammen, neue Schemata zu entwickeln, sowie bisherige weiter zu verfeinern und zu verbessern. Zusätzlich ist es möglich, andere Bereiche neben der eigentlichen Hauptaufgabe speziell zu konfigurieren. Ein kleines Minus sind jedoch die fehlerhaften Schemata, die sich durch diese Konfigurierbarkeit ergeben haben und die Limitationen der transaktionssicheren Datenbank, mit der ein Verzeichnisdienst oft in Verbindung steht.[111]

Skalierbarkeit:

Die Skalierbarkeit ist eine der großen Stärken von LDAP und wird somit folgerichtig mit 3 Punkten als hoch eingestuft. Wie bereits bei der Ausfallsicherheit erwähnt, ist es möglich, die Daten der Verzeichnisse verteilt zu speichern und nahezu beliebig zu replizieren. Dadurch ergibt sich eine offene und effiziente Skalierbarkeit, da man Verzeichnisse beliebig anderweitig abspeichern und verlagern kann. Die einzige Limitation entsteht durch den verbrauchten Speicherplatz der zur Verfügung gestellt werden muss.

Kombinierbarkeit:

Die Kombinierbarkeit ist wieder etwas schwerer zu beurteilen. Auf der einen Seite unterstützen mittlerweile die meisten Programmiersprachen und Betriebssysteme LDAP, auf der anderen Seite gibt es eben solche, die dies noch nicht tun. Auch wenn LDAP vom Prinzip her viele Kombinationsmöglichkeiten, wie zum Beispiel mit dem ACL-Prinzip, hat, sind diese oft noch nicht ausgereift. Da auch die verbundene Datenbank bei OpenLDAP eine Einschränkung darstellt[112], wird eine mittlere Bewertung mit 2 Punkten vorgenommen. Es ist jedoch das Potenzial vorhanden sich in Zukunft deutlich zu steigern.

Performance:

Das letzte Kriterium in der Nutzwertanalyse zur Funktionalität ist die Performance. Hier merkt man deutlich, dass es sich bei LDAP um ein Protokoll mit nur einem Ziel handelt: Schnell Daten aus Verzeichnissen auslesen. Sollte das auf Lesezugriffe optimierte Protokoll dennoch

[110] Vgl. OpenLDAP Foundation (2014)
[111] Vgl. Haberer, P (o.J.)
[112] Vgl. Schwaberow, V. (2001)

nicht genug Performance bieten, um parallele Zugriffe den Anforderungen entsprechend zu verarbeiten, gibt es die Möglichkeit auf die Replikation des gleichen Verzeichnisses zuzugreifen. Insgesamt ist die Performance überzeugend und als hoch einzustufen.[113]

Funktionalität LDAP	*Gewichtung*	*Bewertung*
Funktionsumfang	4,5	2
Konfigurierbarkeit	3	3
Skalierbarkeit	2,5	3
Kombinierbarkeit	3,5	2
Performance	4	3
Gesamt	Faktor 1,5x	66,75

Tabelle 18: Nutzwertanalyse Funktionalität LDAP

Tabelle 18 veranschaulicht die vollständige Nutzwertanalyse von LDAP zu dem Thema Funktionalität. Durch das Einfügen der obengenannten Bewertungen und die Verrechnung mit Gewichtung und Faktor ergibt sich ein Ergebnis von 66,75 Punkten. Dies scheint ein starker Wert zu sein, gerade auch im Verhältnis mit der niedrigeren durchschnittlichen Gewichtung und dem niedrigerem Faktor. Die Funktionalität wird voraussichtlich der Stärkste der 3 Hauptbereiche für LDAP sein.

10.5 Umsetzung

In dem Kapitel der Umsetzung von OpenLDAP ergibt sich das Hauptproblem an den eigentlichen speziellen Anforderungen jedes Unternehmens bzw. jeder Entität die LDAP nutzen will. Die Bewertung für ein weltweit tätiges Unternehmen zur Umsetzung würde ganz anders aussehen, als die Bewertung für einen kleinen Betrieb. Um die Bewertung möglichst relevant für alle Gruppen zu gestalten, wird von einem mittelgroßen Unternehmen ausgegangen. Trotzdem ist anzumerken, dass jedes Unternehmen andere Erwartungen an die Umsetzung stellt und andere Wege wählt.

Support:

Der Support für OpenLDAP ist insgesamt eher undurchsichtig. Auf der offiziellen Homepage wird auf eine offene Mailingliste mit Diskussionen zu OpenLDAP und ein Issue-Tracking System hingewiesen. Außerdem gibt es Verweise auf 3rd Party Supportanbieter mit dem ausdrücklichen Hinweis, dass es keine Vereinbarungen zwischen OpenLDAP und diesen Anbie-

[113] Vgl. Haberer, P (o.J.)

tern gibt.[114] Ein konkretes Problem mit dem Support schnell zu lösen erscheint unrealistisch. Folglich wird der Support mit 1 als niedrig bewertet.

Benutzerfreundlichkeit:

Die Benutzerfreundlichkeit zu bewerten ist erneut stark abhängig vom Benutzer. Da es sich um eine Linux-Umgebung handelt, wird ein entsprechender Nutzer mit grundlegenden Linux und Programmierkenntnissen vorausgesetzt. Diesem sollte es nicht sonderlich schwer fallen, auf einer Kommandozeilenoberfläche zu operieren. Dennoch könnte die verwendete Sprache einige Probleme bereiten und der Benutzer müsste sich die akkurate Nutzung von OpenLDAP erst antrainieren. Eine mittlere Bewertung mit 2 Punkten scheint angemessen.

Implementierungsaufwand:

Der Implementierungsaufwand für ein mittelgroßes Unternehmen ist wahrscheinlich durchschnittlich und folgerichtig mit 2 Punkten zu bewerten. Die Installation von LDAP, samt Datenbank und Server, sollte relativ einfach sein. Es gibt einige Tutorials, die alle nicht zu umfangreich erscheinen. Für ein mittelgroßes Unternehmen wird eine einfache Installation jedoch nicht ausreichend sein. Es ist nicht ersichtlich wie viel Mehraufwand eine Installation in großen Umfang mit sich bringt. Es ist allerdings zu beachten, dass die verschiedenen Schemata alle eingepflegt werden müssen, wenn ein Unternehmen viele verschiedene Verzeichnisse benutzt. Diese müssen fehlerfrei implementiert werden und können nicht mit den Default Einstellungen konfiguriert werden. Dem gegenüber steht das Wissen, dass eine Implementierung in einem Unternehmen immer mit erheblichem Aufwand verbunden ist und es sich hier um keine Ausnahme handeln dürfte. Es bleibt am Ende festzuhalten, dass diese Bewertung spekulativ ist und im Gegensatz zu den Anderen kaum belegt werden kann.

Release Abstände:

Der initiale Release von OpenLDAP 2.4 war am 31.20.2007. Das momentan letzte Release stammt vom 20.09.2014 unter der Version 2.4.40. Das entspricht 40 Versionen auf ca. 6 Jahre und damit mehr oder weniger ein neues Update alle 2 Monate. Die Releasehäufigkeit hat jedoch zunehmend abgenommen. Waren es 2013 noch fünf Versionsupdates, so waren es 2014 nur noch zwei Releases.[115] Dies kann im Zusammenhang stehen mit der Konzentration auf Version 3.0 oder einfach daran liegen, dass die meisten Fehler einfach behoben sind. Die Release-Abstände scheinen in Ordnung und werden als mittel eingestuft.

Dokumentation:

[114] Vgl. OpenLDAP Foundation (2014)
[115] Vgl. ebenda

Die Dokumentation auf der offiziellen Homepage ist sehr umfangreich und kann nur mit 3 Punkten bewertet werden. Die ganzen Dokumente genauer zu betrachten, ist im Umfang und Zeitraum dieser Arbeit nicht möglich, aber Stichproben suggerieren eine sehr ausführliche und verständliche Dokumentation aller Prozesse und Schritte. Dabei werden sogar verschiedene Formate für den Nutzer zur Verfügung gestellt, so wie die Guides zu älteren Versionen.[116]

Umsetzung LDAP	*Gewichtung*	*Bewertung*
Support	3	1
Benutzerfreundlichkeit	3,5	2
Implementierungsaufwand	4	2
Release Abstände	1,5	2
Dokumentation	2,5	3
Gesamt	Faktor 1x	28,5

Tabelle 19: Nutzwertanalyse Umsetzung LDAP

Tabelle 19 stellt die Nutzwertanalyse zur Umsetzung von LDAP dar. Die eingetragenen Werte ergeben zusammen eine Wertung von 28,5 Punkten. Diese scheint auf den ersten Blick deutlich geringer als zuvor, ist aber mit der niedrigen Gewichtung und ohne Faktor als durchaus ordentlich zu erachten.

[116] Vgl. OpenLDAP Foundation (2014)

11 Security Enhanced Linux (SELinux)

Im Folgenden wird SELinux hinsichtlich der Sicherheitsqualität, der Funktionalität und der Umsetzung genauer untersucht und anhand des bereits vorgestellten Bewertungskatalogs ausgewertet.

11.1 SELinux im Überblick

SELinux ist eine Erweiterung des Linux-Kernels und wurde in Kooperation mit der National Security Agency und Red Hat entwickelt. Im Jahre 2000 wurde es unter der GNU GPL veröffentlicht. SELinux beschützt Benutzer und Prozesse, indem es alle Aktionen auf einem System mithilfe der MAC (Mandatory Access Control) überwacht. Beispielsweise wird das Öffnen einer Datei ebenso untersucht, wie der Zugriff auf eine Schnittstelle. Benutzer können zusätzlich die Sicherheitsparameter je nach gewünschter Risikotoleranz konfigurieren.

11.2 Funktionsweise von SELinux

Durch Sicherheits-Labels werden bei SELinux Ressourcen (z.B. Prozesse und Dateien) gekennzeichnet und klassifiziert. Auf diese Weise werden Sicherheitsregeln angewandt, die bestimmen, welcher Benutzer auf welche Ressourcen zugreifen darf. Sicherheits-Labels werden i.d.R. als String aus drei bis vier Worten ausgedrückt. Jedes Wort drückt eine unterschiedliche Komponente des Sicherheits-Labels aus. Die verschiedenen Komponenten sind „user“, „role“, „type“ und „level“ und beziehen sich auf eine Datei oder einen Prozess. Im folgenden Abschnitt werden diese genauer erläutert.

Die erste Komponente ist das „user“ Feld. Dieser Teil des Strings wird gebraucht um Rollen zu gruppieren. Benutzer können verschiedene Rollen zugewiesen bekommen. Typische Rollen sind beispielsweise „user_u“, „system_u“ und „root“. “user_u” ist der Standardwert bei SELinux der Benutzern zugewiesen wird wenn sie sich anmelden. „system_u“ steht für Prozesse, die beim Systemstart ausgeführt wurden. „root“ ist der Wert, der einem Benutzer zugewiesen wird, wenn er sich über die Konsole als „root“ anmeldet. Diese Labels werden beispielsweise benutzt, wenn ein User eine Datei erstellt und diese das Label „user_u“ übernimmt. Wenn Dateien vom System erstellt werden bekommen sie das Label „system_u“.

Die zweite Komponente ist das „role” Feld. Dieses Feld wird benutzt um einer Datei oder einem Prozess eine Sicherheitsgruppe zuzuweisen. Eine Datei hat den Wert „object_r“, wobei Prozesse im Normalfall „system_r“ oder „sysadm_r“ zugewiesen bekommen. Durch die Einteilung in verschiedene Sicherheitsgruppen, können Regeln bestimmt werden, wer ver-

schiedene Prozesse/Dateien ausführen oder aufrufen darf. Diese Methode ist Roles Based Access Control (RBAC) genannt.

Das dritte Feld ist das „type” Feld. Dieses Feld ist wichtig, da es den Subjekttyp festlegt, d.h. es hilft den Sicherheitsregeln zu bestimmen, welche Subjekttypen auf welche Objekttypen zugreifen können. Die letzte Komponente ist das „level” Feld. Diese ist, im Gegensatz zu allen anderen, optional und drückt die Sicherheitsstufe aus.[117]

11.3 Sicherheitsqualität:

Sicherheit:

Ein Hoher Sicherheitsstandard ist durch MAC (Mandatory Access Control) und RBAC (Role Based Access Control) gegeben. Durch den Einsatz von TE (Type Enforcement) wird das Sicherheitslevel weiterhin verstärkt.

Zugriffsrechte:

Die Zugriffsrechteverteilung ist beliebig einstellbar und ist auf alle Dateien und Prozesse anwendbar. Durch die Aufteilung in verschiedene Gruppen mit unterschiedlichen Rechten („user_u“, „system_u“ und „root“) lassen sich die Zugriffsrechte ideal verwalten.

Ausfallsicherheit:

SELinux selbst bietet keinen oder wenig Schutz vor Systemabstürzen, funktioniert im Falle eines Systemabsturzes beim Neustart wie gehabt.[118]

Logging:

SELinux kann in einem Modus ausgeführt werden, indem Sicherheitsverstöße gegen die Sicherheitsrichtlinien protokolliert werden. Des Weiteren werden Aktionen von SELinux in Log-Dateien abgelegt.[119]

Fehlerrate:

Zu diesem Aspekt können keine konkreten Daten/Analysen gefunden werden, jedoch bietet SELinux ein hohes Maß an Sicherheit, weshalb ein hoher Stabilitätsgrad zu erwarten ist.[120]

[117] Vgl. Fedora (2012)
[118] Vgl. Frommel, O. (2005)
[119] Vgl. ebenda
[120] Vgl. NSA (2001), S.4

Aus den obigen Beschreibungen lässt sich somit folgende Bewertung ableiten:

Sicherheitsqualität SELinux	*Gewichtung*	*Bewertung*
Sicherheit	4,5	3
Zugriffsrechte	5	3
Ausfallsicherheit	4	1
Logging	3	3
(Fehlerrate)	3,5	3
Gesamt	Faktor 2x	104

Tabelle 20: Nutzwertanalyse Sicherheitsqualität SELinux

11.4 Funktionalität

Funktionsumfang:

Bezüglich der Funktionalität lassen sich folgende Aspekte zusammenfassen. Klar definierte Richtlinien-Interfaces erlauben es dem Benutzer, die Sicherheitsregeln je nach Wunsch zu konfigurieren. Hauptsächlich bietet SELinux Kontrolle über Prozess-Initialisierungen, Vererbung und Ausführung. Zusätzlich wird die Kontrolle über Dateien und Verzeichnisse gegeben, bzw. die Überwachung von Dateisystemen. Ebenfalls kontrolliert werden Netzwerkanschlüsse, Nachrichten und Netzwerk-Interfaces.[121]

Konfigurierbarkeit:

Sicherheitsregeln und Zugriffsrechte sind je nach Wunsch des Benutzers konfigurierbar. Dies erfordert jedoch ein gewisses Maß an Kenntnissen über das Betriebssystem und die Sicherheitsregeln.

Skalierbarkeit:

Mechanismen, welche die Skalierbarkeit unterstützen sind in SELinux implementiert, jedoch liegen keine konkreten Analysedaten vor.

Kombinierbarkeit:

SELinux kann über zusätzliche Module erweitert werden und bietet die Option, zusätzliche Sicherheitsmodule zu verwenden.[122]

[121] Vgl. Fedora (2008)
[122] Vgl. Gentoo Linux (2015)

Performance:

Die Performance des Linux Kernels unter SELinux wird hierbei überraschenderweise kaum beeinträchtigt, wie es Abbildung 6 beweist.

Table: Macrobenchmark results. The elapsed and system times for a ``time make" on the Linux 2.4.2 kernel sources are shown in minutes and seconds. The latency in seconds and throughput in MBits per second are shown for the WebStone benchmark.

	Base	SELinux	Overhead
elapsed	11:14	11:15	0%
system	00:49	00:51	4%
latency	0.56	0.56	0%
throughput	8.29	8.28	0%

Abb. 6: Macrobenchmark, SELinux Ergebnisse, NSA (Source)

An den Ergebnissen des Makrobenchmarking-Tests der NSA wird deutlich, dass die Performance unter SELinux nur minimal beeinträchtigt wird. Ein präziserer Mikrobenchmarking-Test der NSA untersucht die Performancenachteile und stellt diese in Abbildung 7 dar.[123]

Table: UnixBench system microbenchmarks. File copy throughput is in megabytes per second. The other UnixBench microbenchmarks are in microseconds per loop iteration (or milliseconds for the shell scripts benchmark). These results were converted into units that can be more easily compared with the lmbench results.

Microbenchmark	Base	SELinux	Overhead
file copy 4KB	49.5	48.6	2%
file copy 1KB	40.4	38.6	5%
file copy 256B	23.0	21.0	10%
pipe	6.17	7.17	16%
pipe switching	12.7	15.0	18%
process creation	485	494	2%
execl	2480	2610	5%
shell scripts (8)	659	684	4%

Abb. 7: Mikrobenchmarking unter SELinux, NSA

[123] NSA (o.J.)

An den Ergebnissen des Tests wird deutlich, dass die Performance etwas gesunken ist, jedoch nicht kritisch und in einzelnen Bereichen mehr als in anderen (z.B. pipe und pipe-switching).

Funktionalität SELinux	*Gewichtung*	*Bewertung*
Funktionsumfang	4,5	3
Konfigurierbarkeit	3	3
Skalierbarkeit	2,5	2
Kombinierbarkeit	3,5	2
Performance	4	2
Gesamt	Faktor 1,5x	63,75

Tabelle 21: Nutzwertanalyse Funktionalität SELinux

11.5 Umsetzung

SELinux wurde im Jahre 2000 als Open-Source Projekt von der NSA und Red Hat veröffentlicht und wird seitdem durch die OS-Gemeinschaft unterstützt und weiterentwickelt. Zusätzlich helfen große Organisationen wie beispielsweise Tresys bei der weiteren Entwicklung von SELinux. Es bietet eine flexible Architektur, welche erweitert und nach Belieben konfiguriert werden kann. Wie bereits im Kapitel zur Sicherheitsqualität erklärt wurde, ist die Möglichkeit, verschiedene Sicherheitsstufen (Targeted, MLS, Minimum)[124] einzustellen, ebenfalls vorhanden. Der gesamte Lebenslauf von SELinux und die Gebrauchsanweisung sind ausführlich dokumentiert und für jeden Benutzer jederzeit abrufbar.

Support:

Support ist durch zahlreiche FAQ's, die Open Source Community und durch große Organisationen wie beispielsweise Tresys gegeben.

Benutzerfreundlichkeit:

SELinux ist auf den ersten Blick recht komplex, jedoch kann die Konfigurierung und Benutzung durch GUI's wie z.B. SELinux Administration erleichtert werden.

Implementierungsaufwand:

SELinux ist bereits auf vielen Plattformen implementiert und muss nur noch aktiviert bzw. eingestellt werden. Beispiele für solche Plattformen sind Red Hat Enterprise Linux 4, Fedora und Ubuntu 8.04. Die manuelle Implementierung ist jedoch relativ komplex und aufwändig.[125]

[124] Fedora (2010)
[125] Vgl. NSA (2006), S.7

Release-Abstände:

Neue Inhalte werden in unregelmäßigen Abständen hinzugefügt. Aufgrund der Komplexität von SELinux finden große Updates nicht oft statt, aber stetige Verbesserungen sind dennoch zu erwarten.

Dokumentation:

Die Dokumentation zu SELinux, welche Benutzerhandbücher, Informationen und Veränderungen enthält, ist kostenlos im Internet erhältlich und ausführlich verfasst.

Umsetzung SELinux	*Gewichtung*	*Bewertung*
Support	3	2
Benutzerfreundlichkeit	3,5	2
Implementierungsaufwand	4	3
Release Abstände	1,5	1
Dokumentation	2,5	3
Gesamt	Faktor 1x	34

Tabelle 22: Nutzwertanalyse Umsetzung SELinux

12 Analyse

Im nachfolgenden Kapitel sollen die verschiedenen Auswertungen der unterschiedlichen Lösungen des vorherigen Kapitels zusammengeführt und ausführlich analysiert werden. Dazu werden diese analog zu der bereits geschehenen Auswertung in den Tabellen Sicherheitsqualität, Funktionalität und Umsetzung zusammengefasst und nacheinander vorgestellt. Im Anschluss soll eine Gesamtauswertung basierend auf diesen Ergebnissen erfolgen. Zusätzlich sollen die vorgenommenen Bewertungen ohne den gelegten Fokus (Faktoren) betrachtet werden und eine kritische Betrachtung der Auswertung durchgeführt werden.

12.1 Analyse Sicherheitsqualität

Es soll nun die zusammengeführte Nutzwertanalyse für den Bereich Sicherheitsqualität vorgestellt werden. Dabei ist anzumerken, dass auf Grund von mangelnden Informationen das Kriterium „Fehlerrate" bei 3 der 5 zu bewerteten Lösungen schlecht einsehbar war und dieses Kriterium nicht mit in die Bewertung eingeflossen ist. Zur Vollständigkeit wird es in der untenstehenden Tabelle dennoch aufgeführt.

Sicherheitsqualität	*Gewichtung*	*ExFS/ ACL*	*Samba*	*LXC*	*LDAP*	*SELinux*
Sicherheit	4,5	3	2	2	2 (3)	3
Zugriffsrechte	5	3	3	2	1 (3)	3
Ausfallsicherheit	4	2	1	2	3	1
Logging	3	3	2	1	3	3
~~*(Fehlerrate)*~~	~~3,5~~	~~x~~	~~x~~	~~x~~	~~1~~	~~3~~
Gesamt	Faktor 2x	91	68	60	70	83

Tabelle 23: Nutzwertanalyse Sicherheitsqualität Gesamt

Tabelle 23 zeigt die Nutzwertanalyse zur Sicherheitsqualität mit den 5 zu untersuchenden Lösungen. Die Abkürzungen und Überschriften stehen in dieser Reihenfolge für Extended File Systems/ACL, Samba, Linux-Container, OpenLDAP und SELinux. Wie aus der Tabelle hervorgeht ergibt sich folgendes Ranking:

1. Extended File Systems/ACL **91 Punkte**
2. SELinux **83 Punkte**
3. OpenLDAP **70 Punkte**
4. Samba **68 Punkte**
5. Linux-Container **60 Punkte**

Es ist dabei auffällig, wie viele hohe Bewertungen vergeben werden. Insgesamt neun von 20 möglichen Feldern werden mit der Höchstnote von 3 Punkten bewertet. Ebenso auffällig ist, dass bei den 3 schlechter bewerteten Lösungen in der Wertung zur Sicherheitsqualität jeweils ein Eintrag mit einer 1 vorhanden ist und diese Wertung in nur vier von 20 Feldern verwendet wird. Daraus ergibt sich der Eindruck einer durchgehend mittleren bis hohen Sicherheitsqualität. Unter Betrachtung des Ergebnisses zur Sicherheitsqualität wird jedoch auch deutlich, dass es zwei Gruppen und einen kleinen Bruch gibt. Einerseits Extended File Systems/ACL und SELinux in der Spitzengruppe und dann schwächer OpenLDAP, Samba und Linux-Container. Dieser Eindruck wird durch den anzuwendenden Faktor einer zweifachen Wertung hier zusätzlich verstärkt. Die zwei besser bewerteten Lösungen erreichen dabei beide in den Kategorien „Sicherheit“ und „Zugriffsrechte“ die höchste Bewertungsstufe. Diese Kategorien haben die höchste Gewichtung und tragen so viel zum Gesamtergebnis der Sicherheitsqualität bei. Bei der Ausfallsicherheit ist überraschenderweise OpenLDAP die einzige Lösung, die die Höchstpunktzahl von 3 Punkten erreichen konnte. Im Bereich Logging erreichen Extended File Systems/ACL und SELinux (sowie OpenLDAP) erneut die beste Bewertung. Damit haben diese beiden Lösungen in 3 von 4 Bereichen die Höchstwertung erreicht. Im Fall von SELinux ist die Ausfallsicherheit jedoch problematisch (1 Punkt). Letztlich ist noch anzumerken, dass OpenLDAP die Option hat, mit zukünftig geplanten Implementierungen, eine deutlich positivere Wertung zu erhalten wie aus den Zahlen in den Klammern hervorgeht.

12.2 Analyse Funktionalität

In diesem Kapitel wird die Nutzwertanalyse zu der Funktionalität zusammengefasst und ausgewertet. Der Faktor beträgt 1,5 und befindet sich damit von der Wichtigkeit zwischen Sicherheitsqualität (2) und Umsetzung (1). Die durchschnittlichen Gewichtungen sind dabei ebenso etwas niedriger, als bei der Sicherheitsqualität und etwas höher als bei der Umsetzung.

Funktionalität	Gewichtung	ExFS/ ACL	Samba	LXC	LDAP	SELinux
Funktionsumfang	4,5	1	1	2	2	3
Konfigurierbarkeit	3	3	2	3	3	3
Skalierbarkeit	2,5	2	3	3	3	2
Kombinierbarkeit	3,5	3	3	2	2	2
Performance	4	1	2	3	3	2
Gesamt	Faktor 1,5x	49,5	54,75	66,75	66,75	63,75

Tabelle 24: Nutzwertanalyse Funktionalität Gesamt

Tabelle 24 stellt die Nutzwertanalyse der Funktionalität für die 5 gewählten Lösungen dar. Dabei befinden sich Linux-Container, SELinux und OpenLDAP sehr eng beisammen an der Spitze der Auswertung, während Extended File Systems/ACL und Samba mit ca. 10-15 Punkten weniger beieinander liegen. Insgesamt sieht die Rangfolge wie folgt aus:

1. Linux-Container **66,75 Punkte**
   OpenLDAP **66,75 Punkte**
2. SELinux **63,75 Punkte**
3. Samba **54,75 Punkte**
4. Extended File Systems/ACL **49,5 Punkte**

Die drei funktionaleren Lösungen Linux-Container, SELinux und OpenLDAP haben eine durchgehend mittel bis hohe Bewertung mit je 2-3 mittleren und hohen Bewertungen. Extended File Systems/ACL und Samba sind dabei vor allem in dem Kriterium Funktionsumfang schwach bewertet (1). Auffällig ist, dass es sich bei diesen Bewertungen um die einzigen niedrigen Bewertungen handelt und ebenso, dass die beiden eben genannten Lösungen die Einzigen sind, die bei Kombinierbarkeit die Höchstpunktzahl von 3 Punkten erreichen. Daraus lässt sich erahnen, dass der geringe Funktionsumfang durch die Kombinierbarkeit mit anderen Programmen oder Diensten auszugleichen ist. Trotz dieser Spezialisierung auf eine Funktionalität ist die Performance nicht hoch eingestuft worden. Insgesamt erweckt die Bewertung dennoch den Eindruck, dass alle Lösungen eine gute Funktionalität aufweisen können, auch im Rahmen der gestellten Anforderungen dieser Arbeit.

12.3 Analyse Umsetzung

Als dritte und letzte Analyse soll nun die Umsetzung untersucht werden. Diese ist im Verhältnis zu den anderen Bereichen als weniger wichtig eingestuft worden. Dies liegt an den Anforderungen, die an diese Arbeit gestellt wurden. Das Ziel war die Bestimmung eines sicheren und funktionalen Tools. Dennoch ist die Umsetzung ein imminent wichtiger Punkt und sollte in der Betrachtung nicht vernachlässigt werden. Eine Testimplementierung aller Lösungen im Umfang dieser Ausarbeitung war nicht möglich mit dem gegebenen Zeitfenster.

Umsetzung	*Gewichtung*	*ExFS/*	*Samba*	*LXC*	*LDAP*	*SELinux*

		ACL				
Support	3	2	3	3	1	2
Benutzerfreundlich-keit	3,5	2	3	2	2	2
Implementierungs-aufwand	4	2	1	2	2	3
Release Abstände	1,5	2	3	3	2	1
Dokumentation	2,5	3	3	3	3	3
Gesamt	Faktor 1x	31,5	35,5	36	28,5	34

Tabelle 25: Nutzwertanalyse Umsetzung Gesamt

Tabelle 25 zeigt die gesammelte Nutzwertanalyse der Umsetzung. Dabei befinden sich alle Lösungen innerhalb einer Spanne von 8 Punkten. Die Reihenfolge sieht wie folgt aus:

1. Linux-Container **36 Punkte**
2. Samba **35,5 Punkte**
3. SELinux **34 Punkte**
4. Extended File Systems/ACL **31,5 Punkte**
5. OpenLDAP **28,5 Punkte**

Auffällig ist, dass die Dokumentation mit 3 Punkten durchgehend mit der Bestnote bewertet wurde und es dementsprechend genügend Informationen gibt um eine erfolgreiche Implementierung zu garantieren und etwaige Probleme zu lösen. Im Gegensatz dazu steht der Implementierungsaufwand, der nur bei SELinux mit der Höchstnote bewertet wurde. Trotz guter Dokumentation ist also ein erheblicher Aufwand notwendig für eine erfolgreiche Umsetzung. Die Benutzerfreundlichkeit ist dabei fast immer mittelmäßig und in 4 von 5 Fällen mit der Note 2 bewertet, lediglich Samba bietet die höchste Benutzerfreundlichkeit. Wie bereits aus der Punktspanne zu erkennen ist, gibt es hier 3 stärkere Lösungen, von denen besonders die Linux-Container eine hohe Wertung erhalten. Dennoch gibt es eher nur geringfügige Unterschiede, sodass die Auswahl eher über die anderen beiden Bereiche erfolgen sollte.

12.4 Gesamtanalyse

Im Folgenden soll das Gesamtergebnis aus den Bereichen Sicherheitsqualität, Funktionalität und Umsetzung zusammengefasst und analysiert werden. Zusätzlich wurden die Zeilen „Vergebene Punkte“ und „Gesamt ohne Faktor“ eingefügt, um eine höhere Vergleichbarkeit zu erreichen. „Vergebene Punkte“ ist die Addition der Punkte die insgesamt vergeben worden sind (Ohne Gewichtung) und „Gesamt ohne Faktor“ ist das Ergebnis mit Gewichtung,

aber ohne den Faktor. Die besten Ergebnisse jeder Zeile wurden jeweils farblich hervorgehoben.

Umsetzung	*ExFS/ ACL*	*Samba*	*LXC*	*LDAP*	*SELinux*
Sicherheitsqualität	91	68	60	70	83
Funktionalität	49,5	54,75	66,75	66,75	63,75
Umsetzung	31,5	35,5	36	28,5	34
Vergebene Punkte	32	32	33	32	33
Gesamt ohne Faktor	110	106	110,5	108	118
Gesamt	172	158,25	162,75	165,25	180,75

Tabelle 26: Zusammenfassung der Ergebnisse der Nutzwertanalysen

Tabelle 26 veranschaulicht die kombinierten Ergebnisse der verschiedenen Analysen. Dabei hat SELinux die Höchstpunktzahl erreicht. Die gesamten Lösungen befinden sich in einem Rahmen von circa 20 Punkten. Geordnet sieht die Rangliste wie folgt aus:

1. SELinux **180,75 Punkte**
2. Extended File Systems/ACL **172 Punkte**
3. OpenLDAP **165,25 Punkte**
4. Linux-Container **162,75 Punkte**
5. Samba **158,25 Punkte**

Die Höchstpunktzahl resultiert dabei aus einer durchgehend hohen Bewertung in allen drei Bereichen. Auffällig ist dabei, dass SELinux in keinem Bereich am meisten Punkte erreicht hat, sondern dies Extended File Systems/ACL (Sicherheitsqualität), Linux-Container (Funktionalität & Umsetzung) und LDAP (Funktionalität) gelungen ist. SELinux und Samba sind damit die einzigen Lösungen ohne Höchstwertung, befinden sich aber an unterschiedlichen Enden der Rangliste. Bei den anderen drei Lösungen von einer wirklichen Spezialisierung zu sprechen ist jedoch auch nicht angebracht. Insgesamt liegen die Wertungen überall sehr eng zusammen. Dies wird durch die vergebenen Punkte belegt, die sich in einer Spanne von 32 bis 33 Punkten befindet und zeigen, dass die Bewertung basierend auf der Analyse der einzelnen Verfasser auf einem ähnlichen Niveau abgelaufen ist. Insgesamt befindet sich die durchschnittliche Bewertung pro Kriterium knapp über 2 Punkten und ist damit als zufriedenstellend einzuordnen. Lediglich die unterschiedlichen Gewichtungen und Faktoren sorgen für eine Diskrepanz zwischen den verschiedenen Lösungen. Dies ist durch die Schwerpunkte der jeweiligen Lösungen zu erklären und ein erwartetes Ergebnis. Dabei ist der gesetzte Faktor von nicht zu vernachlässigender Bedeutung. Wie die Zeile Gesamt ohne Faktor ver-

anschaulicht verändert sich das Ergebnis ohne Faktoren dahingehend, dass die Linux-Container von Platz 4 auf Platz 2 vorrücken und Extended File Systems/ACL und LDAP überholen. Dies ist zu begründen durch die Schwäche der Linux-Container im Bereich Sicherheitsqualität.

13 Lösungsansätze

Für die weitere Betrachtung ist es enorm wichtig, das Zusammenspiel zwischen den einzelnen Lösungen zu verstehen und auszuwerten, um ein möglichst ideales Gesamtergebnis erzielen zu können.

Dazu sollen die Lösungsansätze in einer Gesamtrelation dargestellt werden. Hierzu wird mit der untersten Ebene, dem Extended File Systems, als Basis begonnen, um anschließend den stufenweisen Aufbau darzustellen.

Das Extended File System in Kombination mit ACL ist bereits in der Lage die Grundanforderung der Zugriffsverwaltung abzudecken. Dabei werden keine weitergreifenden Prozesse oder Funktionen bereitgestellt. Um diese Funktionalitätserweiterung einzuführen wird als nächster Baustein OpenLDAP auf dem Extended File System eingerichtet. Dieser Zuwachs bietet den enormen Vorteil der flexiblen Skalierbarkeit sowie der verteilten Ansprache von Speichersystemen. Negativ sind jedoch weiterhin die unzureichende Benutzerfreundlichkeit von OpenLDAP und die vorherrschende Unsicherheit des Linux Kernels des hier vorliegenden Konzepts zu bewerten.

Die Benutzerfreundlichkeit kann mit der Erweiterung durch das Tool Samba gesteigert werden, dadurch wird für den Benutzer eine bekannte Windows Oberfläche emuliert. Um zusätzlich die Sicherheitsqualität zu steigern, wird der Einsatz von SELinux empfohlen. SELinux bietet den Vorteil des gehärteten Kernels, wodurch weitreichende Sicherheitsmechanismen hinzugefügt werden. Dieser soeben dargestellte Lösungsansatz ist zusätzlich in Abbildung 8 gezeigt.

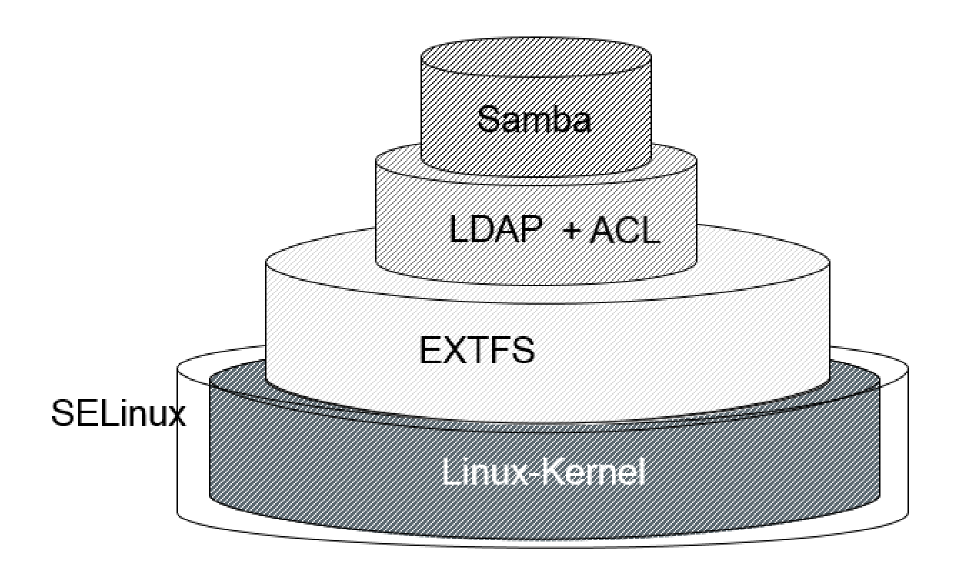


Abb. 8: Erster Lösungsansatz

Allgemein ist jedoch davon auszugehen, dass eine Steigerung der Kombinationen unweigerlich zu einer Komplexitätssteigerung führt. Dieser nicht zwingend homogene Ansatz kann ungewollte Fehlerquellen und einen erhöhten Implementierungsaufwand herbeiführen.

In einem letzten Schritt kann dieses Konzept oder Teile des Konzepts in eine LXC-Umgebung eingebunden werden. Hierdurch kann zusätzlich die Zugriffsicherheit erhöht sowie durch die gesonderte Gruppierung spezifischer Prozesse eine hoch-performante Lösung erzielt werden. Eine Erweiterung nach Benutzervorstellungen ist problemlos anwendbar durch die Klonfunktion. Diese generiert einen in der Konfiguration identischen Container, der zusätzlich spezifisch modifiziert werden kann. Diese Gesamtkombination wird in Abbildung 9 veranschaulicht.

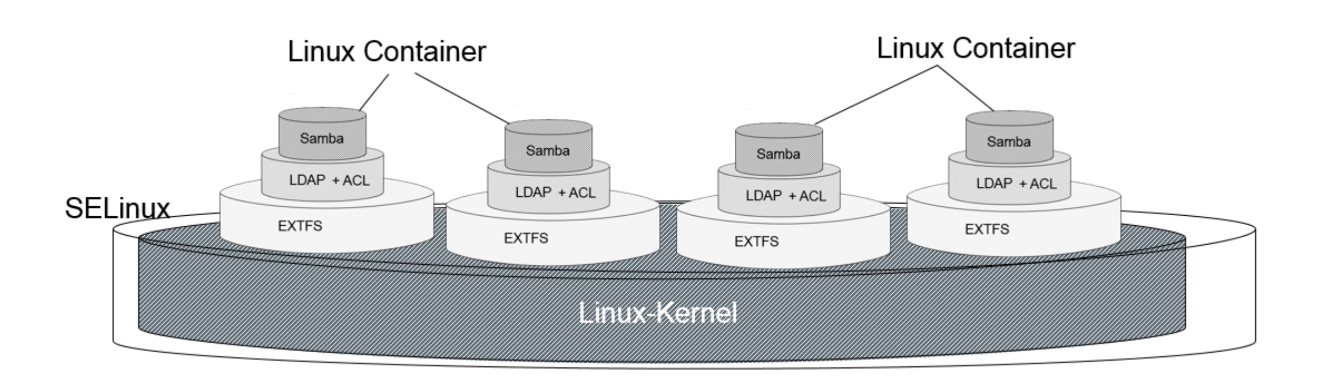


Abb. 9: Zweiter Lösungsansatz

Es bleibt festzuhalten, dass dieser Ansatz höchste Anforderungen an das Implementierungsteam stellt und somit ein erhöhtes Risiko beherbergt. Zudem sind zu diesem Zeitpunkt keine Referenzkunden vorzuweisen, die eine reine LXC-Umgebung betreiben. Dies stellt die Verlässlichkeit stark in Frage und sorgt dafür das Linux Container nicht ohne tiefgreifende Betrachtung empfohlen werden kann.

Abschließend lässt sich festhalten, dass sich alle Lösungen im Anforderungsrahmen befinden und für eine Implementierung in Frage kommen.

Im Verlaufe des Projekts hat sich das in Abbildung 8 vorgestellte Modell als eine verlässliche und praktikable Lösung herausgestellt. Aus diesem Grund wird dieser Lösungsansatz von den Verfassern empfohlen. Ist der Anwender gewillt einen Mehraufwand in Bezug auf Implementierungsaufwand im Gegenzug für eine höhere Sicherheit und maximale Performance einzugehen, wäre auch das in Abbildung 9 vorgestellte Modell eine denkbare Lösung. Ist das Ziel lediglich ein grundlegendes System aufzubauen, sollte die Kombination aus Extended File System und LDAP sowie eventuell Samba gewählt werden.

14 Zusammenfassung

Zusammenfassend lässt sich sagen, dass ein allgemeingültiger Lösungsansatz nicht existiert. Dieses Ergebnis wird durch die vorhergehende Analyse verdeutlicht, da kein Sicherheitsmechanismus gegenüber den Anderen bezüglich jedes einzelnen Kriteriums überzeugt. Nach Auswertung des Kriterienkatalogs wurde zudem deutlich, dass der Mechanismus mit den meisten Punkten, in keiner Rubrik mit dem ersten Platz abgeschnitten hat, was die Vermutung unterstützt, dass es sich hierbei nur um ein Kompromissprodukt handeln könnte.

Somit kann die Forschungsfrage, welcher Sicherheitsmechanismus den besten Lösungsansatz für die Verwaltung und Realisierung der Zugriffskontrolle sowie der Abschottung verschiedener File-Systeme ermöglicht, mit dem oben genannten Resultat nicht reinen Gewissens beantwortet werden. Vielmehr muss nach einer Kombination aus den besten Sicherheitsmechanismen gesucht werden, um die gestellte Forschungsfrage im vollsten Umfang zu beantworten.

Nach Betrachtung der fünf unabhängigen Sicherheitsmechanismen deuten sich unterschiedliche Stärken jedes einzelnen Mechanismus an. Daher erfolgt im Folgenden keine Empfehlung eines einzelnen Sicherheitsmechanismus, sondern der Kombination mehrerer Sicherheitsmechanismen, die ihre Stärken in unterschiedlichen Bereichen erfüllen.

Die von den Verfassern empfohlene Lösung für den vorliegenden Anwendungsfall kann dem Kapitel 13 entnommen werden. Diese zeichnet sich durch die Vereinbarung von hoher Sicherheitsqualität und voller Funktionalität bei einem akzeptablen Umsetzungsaufwand aus. Für andere Schwerpunkte wurden jeweils alternative Lösungsvorschläge aufgezeigt.

Das in der Einleitung genannte Ziel, der Durchführung einer Marktstudie anhand eines Kriterienkatalogs mit dem Ziel unterschiedliche Sicherheitsmechanismen der Zugriffsverwaltung zu beleuchten, um einen Lösungsvorschlag zu erarbeiten, wurde erfüllt.

Während der Beleuchtung der einzelnen Sicherheitsmechanismen hat sich zudem ein sehr interessanter Zukunftstrend angedeutet. Fast alle Mechanismen weisen eine hohe Update- und Innovationsrate auf, sodass auch in naher Zukunft auf diesem Gebiet mehr oder weniger große positive Veränderungen zu erwarten sind, die den hier angesprochenen Anwendungsfall gegebenenfalls noch besser unterstützen können. Zudem ist ein steigendes Wachstum in der Open Source Gemeinde zu vermelden, die das Angebot an innovativen Open Source Lösungen vervielfältigen wird. Die Entscheidung der .Versicherung, diesem Trend zu folgen, ist als wegweisend zu betrachten.

Anhang

1. Free Redistribution	The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.
2. Source Code	The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well-publicized means of obtaining the source code for no more than a reasonable reproduction cost preferably, downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed.
3. Derived Works	The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.
4. Integrity of The Author's Source Code	The license may restrict source-code from being distributed in modified form only if the license allows the distribution of "patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modified source code. The license may require derived works to carry a different name or version number from the original software.
5. No Discrimination Against Persons or Groups	The license must not discriminate against any person or group of persons.
6. No Discrimination Against Fields of Endeavor	The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research.
7. Distribution of License	The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.
8. License Must Not Be Specific to a Product	The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.
9. License Must Not Restrict Other Software	The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open-source software.
10. License Must Be Technology-Neutral	No provision of the license may be predicated on any individual technology or style of interface.

Kriterien zur Bestimmung eines Open Source Produktes

Quellenverzeichnisse

Archlinux (2015): Linux Containers, https://wiki.archlinux.org/index.php/Linux_Containers, Abruf: 11.01.2015.

Banning, Jens (2001): LDAP unter Linux, Netzwerkinformationen in Verzeichnisdiensten verwalten (Open Source Library), 2. Auflage, München: Addison-Wesley.

Anonymisiert (2014): .Versicherung, Stuttgart, persönliches Gespräch am 17.12.2014.

Cisco (2014): Linux Containers, http://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/dc-partner-red-hat/linux-containers-white-paper-cisco-red-hat.pdf, Abruf: 13.01.2015.

Clark, Jack (2014): Google: „Everything at Google runs in a container", http://www.theregister.co.uk/2014/05/23/google_containerization_two_billion/, Abruf: 12.01.2015.

Debian (2011): Evading from Linux
ers, http://web.archive.org/web/20140109184419/http:/blog.bofh.it/debian/id_413, Abruf: 13.01.2015

Docker (2014): Use Cases, https://www.docker.com/resources/usecases/, Abruf: 10.01.2015.

Dundee, Paul / R., Jelmer / H., John (o.J.): Samba 3 - HOWTO, https://www.samba.org/samba/docs/man/Samba3-HOWTO/speed.html, Abruf: 13.01.2015.

Eckstein, Robert / Collier-Brown, David / Kelly, Peter (1999): Using Samba, 1. Auflage, Sebastopol: O'Reilly Media.

Eggeling, Thorsten (2014): Netzwerken mit Samba - so geht's, http://www.pcwelt.de/ratgeber/Netzwerken_mit_Samba_-_so_geht_s-Linux-8530128.html, Abruf: 13.01.2015.

Fedora (2008): SELinux/FC5Features, http://fedoraproject.org/wiki/SELinux/FC5Features, Abruf: 15.01.2015.

Fedora (2010): SELinux/Policies, https://fedoraproject.org/wiki/SELinux/Policies, Abruf: 15.01.2015.

Fedora (2012): Security context, https://fedoraproject.org/wiki/Security_context?rd=SELinux/SecurityContext, Abruf: 15.01.2015.

Fenzi, Kevin / Wreski, Dave (2004): Linux Security HOWTO, http://www.tldp.org/HOWTO/Security-HOWTO/, Abruf: 11.01.2015.

Free Software Foundation (2014): GNU operating system, http://www.gnu.org/licenses/quick-guide-gplv3.html, Abruf: 13.01.2015.

Frommel, Oliver (2005): Fedora Core 4, https://www.linux-user.de/ausgabe/2005/08/083-fedora/index.html, Abruf: 15.01.2015.

Gentoo Linux (2015): SELinux/Tutorials/Creating your own policy module file, http://wiki.gentoo.org/wiki/SELinux/Tutorials/Creating_your_own_policy_module_file, Abruf: 15.01.2015.

Golembowska, Anne u.a. (2012): Entwicklung eines Modells zur Bewertung von Open Source Produkten hinsichtlich eines produktiven EinsatzesSeminararbeit, Stuttgart:

Gollub, Daniel / Seyfried, Stefan (2010): Ressourcen-Verwaltung mit Control Groups (cgroups), http://www.pro-linux.de/artikel/2/1464/ressourcen-verwaltung-mit-control-groups-cgroups.html, Abruf: 11.01.2015.

Graber, Stéphane (2014): LXC 1.0, https://www.stgraber.org/2014/01/01/lxc-1-0-security-features/, Abruf: 14.01.2015.

Grimmer, Lenz (2013): Linux Containers explained.

H., John (o.J.): Samba - HOWTO Collection, https://www.samba.org/samba/docs/man/Samba-HOWTO-Collection/IntroSMB.html, Abruf: 13.01.2015.

Haberer, Petra (o.J.): LDAP verstehen, http://www.mitlinx.de/ldap/, Abruf: 12.01.2015.

Heise (1998): Netscape veröffentlicht Navigator-Quelltext, http://www.heise.de/newsticker/meldung/Netscape-veroeffentlicht-Navigator-Quelltext-11625.html, Abruf: 03.01.2015.

Ibáñez, Roger Ferrer (2014): What is Eiciel?, http://rofi.roger-ferrer.org/eiciel/?s=2, Abruf: 14.01.2015.

IBM (2012a): Shelter Mutual Insurance Company slashes costs and complexity, http://www-03.ibm.com/software/businesscasestudies/us/en/corp?synkey=X945885V97903L38, Abruf: 10.01.2015.

IBM (2012b): EFiS EDI Finance Service AG boosts business flexibility and efficiency, http://www-03.ibm.com/software/businesscasestudies/us/en/corp?synkey=D370982Q60606T62, Abruf: 10.01.2015.

IBM (2012c): Cryptographic advances for Linux on System z Applications, http://www-01.ibm.com/common/ssi/cgi-http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=PM&subtype=SP&appname=STGE_ZS_ZS_USEN&htmlfid=ZSS03052USEN&attachment=ZSS03052USEN.PDF#loaded, Abruf: 10.01.2015.

IBM (2013a): Linux on IBM z Systems, http://www-03.ibm.com/systems/z/os/linux/about.html, Abruf: 10.01.2015.

IBM (2013b): Security, http://www-03.ibm.com/systems/z/os/linux/solutions/security.html, Abruf: 10.01.2015.

Kerrisk, Michael (2015): Namespaces, http://man7.org/linux/man-pages/man7/namespaces.7.htm, Abruf: 11.01.2015.

Kleikamp, Dave (2013): JFS for Linux, http://jfs.sourceforge.net/, Abruf: 14.01.2015.

Kofler, Michael (2008): Linux: Installation, Konfiguration, Anwendung, 8. Auflage, München: Addison-Wesley.

Lang, Marco (2012): Open Source von A-Z, http://www.gizlog.de/2012/open-source-von-a-z/#zh-3, Abruf: 13.01.2015.

Linux Containers (o.J.): What's LXC?, https://linuxcontainers.org/lxc/introduction/, Abruf: 05.01.2015.

Linux Kernel Organization (2014a): Extended Filesystem EXT4, https://ext4.wiki.kernel.org/index.php/Main_Page, Abruf: 14.01.2015.

Linux Kernel Organization (2014b): Extended Filesystem EXT4 Releases, https://ext4.wiki.kernel.org/index.php/Ext4:News, Abruf: 14.01.2015.

Linuxtopia (o.J.a): Linuxtopia, http://www.linuxtopia.org/LinuxSecurity/LinuxSecurity_Introduction1.html, Abruf: 11.01.2015.

Linuxtopia (o.J.b): Linuxtopia, http://www.linuxtopia.org/LinuxSecurity/LinuxSecurity_Introduction_Security.html, Abruf: 11.01.2015.

LinuxWiki (2011): Chroot, http://linuxwiki.de/chroot, Abruf: 13.01.2015.

LinuxWiki (2013): Extended Filesystem EXT3, http://linuxwiki.de/ext3, Abruf: 14.01.2015.

Lowe, Scott (2013): A Brief Introduction to Linux Containers, http://blog.scottlowe.org/2013/11/25/a-brief-introduction-to-linux-containers-with-lxc/, Abruf: 10.01.2015.

Marmol, Victor / Jnagal, Rohit (2013): Let me contain that for you, http://www.linuxplumbersconf.org/2013/ocw/system/presentations/1239/original/lmctfy%20(1).pdf, Abruf: 12.01.2015.

Novell (2011): Access Control Lists in Linux, https://doc.opensuse.org/documentation/html/openSUSE_121/opensuse-security/cha.security.acls.html, Abruf: 01.02.2015

Noyes, Katherine (2010): Why Linux Is More Secure Than Windows, http://www.pcworld.com/article/202452/why_linux_is_more_secure_than_windows.html, Abruf: 11.01.2015.

NSA (o.J.): UnixBench, https://www.nsa.gov/research/_files/selinux/papers/freenix01/node15.shtml, Abruf: 15.01.2015.

NSA (2001): Integrating Flexible Support for Security Policies into the Linux Operating System, https://www.nsa.gov/research/_files/selinux/papers/slinux.pdf, Abruf: 15.01.2015.

NSA (2006): Implementing SELinux as a Linux Security Module, https://www.nsa.gov/research/_files/publications/implementing_selinux.pdf, Abruf: 15.01.2015.

Open Source Initiative (o.J.a): The Open Source Definition, http://opensource.org/osd, Abruf: 03.01.2015.

Open Source Initiative (o.J.b): The BSD 2-Clause License, http://opensource.org/licenses/BSD-2-Clause, Abruf: 03.01.2015.

OpenLDAP Foundation (2014): OpenLDAP, http://www.openldap.org/, Abruf: 14.01.2015.

Prakasha, Swayam (o.J.): Security in Linux, http://www.linuxuser.co.uk/features/security-in-linux, Abruf: 11.01.2015.

Samba (o.J.a): Opening Windows to a Wider World, https://www.samba.org/samba/, Abruf: 13.01.2015.

Samba (o.J.b): Previous Release Announcements, https://www.samba.org/samba/history/, Abruf: 13.01.2015.

Samba (o.J.c): Samba Support, https://www.samba.org/samba/support/, Abruf: 13.01.2015.

Sarton, J. J. (o.J.): Rechte unter Unix Systemen, http://www.knilse.de/download/acl.html, Abruf: 14.01.2015.

Schmidt, Meik (2006): Samba einrichten und Linux als Fileserver, http://www.pc-erfahrung.de/linux/linux-samba.html, Abruf: 13.01.2015.

Schnabel, Patrick (2014): LDAP – Lightweight Directory Access Protocoll, www.elektronik-kompendium.de/sites/net/0905021.htm, Abruf: 12.01.2015.

Schwaberow, Volker (2001): OpenLDAP Praxis, Straffe Verwaltung, in: Linux Magazin, 05/2001, http://www.linux-magazin.de/Ausgaben/2001/05/Straffe-Verwaltung.

Simon, Christian (2014): Härtungs- und Sicherheitskonzepte für Web- und Applikationsser-verMasterarbeit, München:

Suse (2003): POSIX ACL, http://users.suse.com/~agruen/acl/linux-acls/online/, Abruf: 14.01.2015.

Suse (2011): Wichtige Dateisysteme in Linux, https://www.suse.com/de-de/documentation/sles10/book_sle_reference/data/sec.filesystems.major.html, Abruf: 14.01.2015.

Tobler, Michael (2001): Inside Linux, o.O.: New Riders.

Ubuntu (2014): LXC, https://help.ubuntu.com/lts/serverguide/lxc.html, Abruf: 13.01.2015.

Sarathy, Bhavny / Walsh, Dan (2013): RedHat Summit: Linux Containers Overview & Roadmap, http://rhsummit.files.wordpress.com/2013/06/sarathy_w_0340_secure_linux_containers_roadmap.pdf, Abruf: 10.01.2015.

Wikipedia (2014): GNU Lesser General Public License, http://de.wikipedia.org/wiki/GNU_Lesser_General_Public_License, Abruf: 11.01.2015.

Winter, Stefan (2014): Management von Lieferanteninnovationen, Eine gestaltungsorientierte Untersuchung über das Einbringen und die Bewertung, Wiesbaden: Springer.

Xavier, Miguel G. u.a. (o.J.): Performance Evaluation of Container-based Virtualization for High Performance Computing Environments, http://marceloneves.org/papers/pdp2013-containers.pdf, Abruf: 11.01.2015.

ZyTrax Inc. (2015): LDAP for Rocket Scientists, http://www.zytrax.com/books/ldap/ch6/, Abruf: 13.01.2015.

z/Journal (2008): Taking the Secure Migration Path to IT Virtualization, in: z/Journal, 08/2008, http://www.vm.ibm.com/devpages/spera/zJAugu08.pdf.